Title
Randomization Tests under the Potential Outcomes Framework

Author
Wu, Jason

Publication Date
2019

Peer reviewed|Thesis/dissertation

Randomization Tests under the Potential Outcomes Framework

By

Jason Wu

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Peng Ding, Chair
Professor Avi Feller
Professor Adityanand Guntuboyina

Summer 2019

Randomization Tests under the Potential Outcomes Framework

Abstract

Randomization Tests under the Potential Outcomes Framework

by

Jason Wu

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Peng Ding, Chair

The Fisher randomization test (FRT) is appropriate for any test statistic, under a sharp null hypothesis that can recover all missing potential outcomes. However, it is often of interest to test a weak null hypothesis that the treatment does not affect the units on average. To use the FRT for a weak null hypothesis, we must address two issues. First, we need to impute the missing potential outcomes although the weak null hypothesis cannot determine all of them. Second, we need to choose a proper test statistic. For a general weak null hypothesis, we propose an approach to imputing missing potential outcomes under a compatible sharp null hypothesis. With this imputation scheme, we advocate a studentized statistic. The resulting FRT has multiple desirable features. First, it is model-free. Second, it is finite-sample exact under the sharp null hypothesis that we use to impute the potential outcomes. Third, it conservatively controls large-sample type I errors under the weak null hypothesis of interest. Therefore, our FRT is agnostic to treatment effect heterogeneity. We establish a unified theory for general factorial experiments. We also extend it to stratified and clustered experiments.

Contents

Contents **i**

List of Figures **iii**

List of Tables **v**

1 Introduction **1**
1.1 Literature Review 1
1.2 Our Contributions 2
1.3 Notation 3

2 Framework for Randomization Tests **4**
2.1 Completely Randomized Experiments 4
2.2 Fisher Randomization Tests 4
2.3 Asymptotics for Finite Population Inference 6

3 Test Statistics **8**
3.1 Proper Test Statistics 8
3.2 Studentized statistic 8
3.3 Box-Type Statistic 9
3.4 Statistics from Ordinary Least Squares 10

4 Special Cases **12**
4.1 One-Way Analysis of Variance 12
4.2 Treatment-Control Experiments 12
4.3 Trend Tests 13
4.4 Binary Outcomes 13
4.5 2^K Factorial Designs 14
4.6 Hodges–Lehmann Estimation 15
4.7 Testing Inequalities 15

4.8 Cluster-Randomized Experiments 16

5 Extensions **18**
5.1 Stratified Randomized Experiments 18
5.2 Multiple Outcomes and Multiple Testings 20

6 Simulations **23**
6.1 Breaking the Box-Type Statistic 23
6.2 Confidence Regions 25
6.3 A treatment-control simulation 27

7 Applications and Discussion **31**
7.1 Financial Incentives for Exercise 31
7.2 A 2^2 Factorial Experiment for Grades 32
7.3 Discussion 33

Bibliography **34**

A Proofs and Lemmas for the Main Text **40**
A.1 Technical Lemmas 40
A.2 Proofs of Main Text Results 45

B Extra background material **56**
B.1 Proofs of other results 56
B.2 Computational Details of the Simulations 62
B.3 More about treatment-control 67
B.4 More on linear models and Huber–White estimation 72
B.5 More on the one-way layout 77
B.6 More on the Box-type and Wald-type statistics 78
B.7 More on vector potential outcomes 83
B.8 Technical matters for FRT 84

List of Figures

6.1 Histograms of FRT p-values under various settings and sample sizes. Gray bars indicate p-values from a F statistic, while transparent bars indicate p-values from the X^2 statistic. We display smaller p-values with a finer resolution because most hypothesis tests are conducted at levels close to 0. A dashed line indicating the Unif$(0, 1)$ density is added for reference purposes. 24

6.2 For τ_1 and τ_2 individually, the FRT and asymptotic approximation give nearly identical confidence intervals. For the second main effect, the FRT confidence interval is shifted due to the discrete resolution. 26

6.3 The left graph shows the FRT confidence region is again close to its asymptotic approximation, but the former is noticeably larger. The right graph is a scatter plot of p-values from testing $\tau_1 = \tau_2 = 0$ repeatedly from the original set of potential outcomes, zooming in on the region where they are less than 0.1. 26

6.4 For N=20, the 0 to 0.01 bin density is 9.55 for the χ^2 approximation and 4.8 for the FRT. The 0.02 to 0.03 density is 2.75 for the FRT. For small samples, the tests do not perform as their asymptotics suggest. 27

6.5 A scatter plot of p-values from the FRT versus those from the χ^2 approximation, when N=20. The dashed line on which the two p-values would be equal is included for reference. 28

6.6 Two realizations of the permutation distribution, under different treatment assignments for potential outcomes B and N=20. In the first case, the p-value using the χ^2 approximation was < 0.01 while the p-value using the FRT was > 0.052. In the latter case, the p-values calculated from either method were both near 0.01. 30

B.1 For our simulated data, the asymptotic 0.95 joint confidence region for τ_1, τ_2. The vertical and horizontal lines are the endpoints of the 0.95 CI for τ_1, τ_2, respectively. The point $(\hat{\tau}_1,\hat{\tau}_2)$ is also shown. The entire plotting region corresponds to our search region. 66

List of Tables

7.1 Analyzing [17]'s data. We report p-values as percents, and calculate the FRT p-values using 10^4 Monte Carlo simulations. 32
7.2 Analyzing [1]'s data. We report p-values as percents, and calculate the FRT p-values using 10^4 Monte Carlo simulations. 33

Acknowledgments

We thank the Statistics Department staff at UC Berkeley, such as La Shana Porlaris, Mary Melinn, Deb Nolan, Erin Blanton, and Laura Slakey, among others. Their tireless efforts have made possible a supportive academic community. We thank the faculty in the Statistics Department, such as Peng Ding, Avi Feller, Sam Pimentel, Aditya Guntuboyina, Deb Nolan, and Hank Ibser. They have provided excellent research, teaching, and personal advice. We thank all the graduate students in the Statistics Department. They fostered a warm and welcoming environment, and always made time to show genuine concern, despite their busy schedules. We thank our parents who have nurtured us and provided generous financial backing, both before and during college. We thank the Associate Editor, two reviewers, Guilaume Basse, Joel Middleton, and Zach Branson for helpful comments. We gratefully acknowledge financial support from the National Science Foundation (DMS RTG # 1745640 for Jason Wu; DMS grant # 1713152 for Peng Ding.

Chapter 1

Introduction

This dissertation is organized as follows. The initial chapters serve as the main text. The first appendix contains proofs for results in the main text, along with the needed auxiliary results. The second appendix contains additional background information that can supplement the material in the main text and the first appendix.

1.1 Literature Review

Randomization is the cornerstone of statistical causal inference [30, Section II]. It creates comparable treatment groups on average. More fundamentally, it justifies the Fisher randomization test (FRT). Under Fisher's sharp null hypothesis, the treatment does not affect any units whatsoever, and the distribution of any test statistic is known over all randomizations [30, 86, 83, 46]. Therefore, the FRT delivers a finite-sample exact p-value. In fact, many parametric and non-parametric tests are approximations to the FRT [27, 77, 52, 13, 21, 14, 55].

Another formulation of the FRT relies on exchangeability of outcomes under different treatments [77, 42, 79]. They called this formulation a "permutation test". [53] accentuated the importance of the treatment assignment mechanism to justify the FRT, without assuming that the outcomes are exchangeable. [86] extended the FRT using [72]'s potential outcomes. He defined a null hypothesis to be sharp if it can determine all missing potential outcomes. One of his insights was that any test statistic has a known distribution under a sharp null hypothesis, and therefore the FRT is finite-sample exact.

Randomized experiments are increasingly popular in the social sciences [26, 37, 46, 5]. In such applications, testing sharp null hypotheses may not answer the questions of interest. Researchers often want to test weak null hypotheses that the treatment has zero effects on average. The ideal testing procedure must leave room for treatment effect heterogeneity. Unfortunately, weak null hypotheses cannot determine all missing potential outcomes, even though the distributions of test statistics depend on them in general. Consequently, simple FRTs may not be directly applicable for testing weak null hypotheses.

Testing weak null hypotheses with the FRT is a delicate matter. Although sometimes we can still use the same FRTs, we need to modify the interpretations without sharp null

hypotheses [84, 81, 82, 16]. Not all FRTs can preserve type I errors for weak null hypotheses even asymptotically. The famous Neyman–Fisher controversy ties into this issue for randomized block designs and Latin square designs [73, 88]. [36] and [60] gave empirical evidence from simulations, and [24] gave a theoretical analysis of the one-way layout. Two strategies exist for using FRTs to test weak null hypotheses. The first strategy realizes that weak null hypotheses become sharp given appropriate nuisance parameters. It maximizes the p-values over all values of the nuisance parameters or their confidence sets [74, 78, 57, 25]. However, it can be computationally burdensome and lacks power when the nuisance parameters are high dimensional. The second strategy uses conditional FRTs. It relies on partitioning the space of all randomizations, and in some subspaces, certain test statistics have known distributions under the weak null hypotheses [4, 7]. It can be restrictive and is not applicable in general settings.

1.2 Our Contributions

We propose a strategy for testing a general hypothesis in a completely randomized factorial experiment. The null hypothesis asserts that certain average factorial effects are zero. It is therefore weak and cannot determine all missing potential outcomes. Our strategy has two components.

First, we specify a sharp null hypothesis. It must imply the weak null hypothesis of interest and be compatible with the observed data. Treatment-unit additivity holds under this sharp null hypothesis. In particular, it implies constant factorial effects of and beyond the weak null hypothesis. Under this sharp null hypothesis, we can impute all missing potential outcomes.

Second, we use the FRT with a studentized test statistic. Like other test statistics, its sampling distribution depends on unknown potential outcomes in general. Thus, this distribution is outside our grasp. Fortunately, the FRT generates a proxy distribution under the above sharp null hypothesis. This proxy distribution stochastically dominates the unknown one asymptotically. The stochastic dominance relationship between them enables us to construct an asymptotically conservative test. Therefore, for testing the weak null hypothesis, we recommend the FRT with the studentized statistic. Without studentization, the FRT may not control type I error even asymptotically. We examine several existing test statistics that exhibit this unwanted behavior.

The idea of studentization already surfaces in the literature. [71], [48], [49], [50] and [19] conducted permutation tests with studentization. These tests assumed that the outcomes are independent draws. In our formulation, the random treatment assignment drives the statistical inference with fixed potential outcomes. We do not assume any exchangeability of outcomes. In this special setting, our theory transmits many new features. First, the sampling distribution of the studentized statistic is not asymptotically pivotal. Rather, the approximate distribution generated by the FRT is. This is distinct from the iid samples setting, where the former distribution is itself asymptotically pivotal. Second, the FRT is conservative for the weak null hypothesis. This aspect of finite-population causal inference

[72, 46, 24] was absent in the literature on permutation tests. Third, studentization helped [6] and [38] achieve better second order accuracy in the bootstrap. In contrast, we use it for better first order accuracy, i.e., to control asymptotic type I error. The bootstrap is another resampling method for testing weak null hypotheses. In relation to the bootstrap, FRTs have an additional advantage of being finite-sample exact under sharp null hypotheses. Although the bootstrap has been a workhorse for many other statistical problems, [45] recently fused its ideas with finite population causal inference.

1.3 Notation

We summarize the notations to be used throughout the manuscript. Let 1_n and 0_n be vectors of n 1's and 0's, respectively. Let $1(\cdot)$ denote the indicator that an event happens. Let $A \succeq 0$ and $A \succ 0$ if A is positive semi-definite and positive definite, respectively. Write $A \succeq B$ if $A - B \succeq 0$. For a diagonalizable matrix A, let $\lambda_j(A)$ be its j-th largest eigenvalue. Let $\text{diag}\{\cdot\}$ be a diagonal or block-diagonal matrix. If (X_N) is a sequence of random variables indexed by N, write $X_N \xrightarrow{\text{d}} X$, $X_N \xrightarrow{P} X$, $X_n \xrightarrow{\text{as}} X$ for convergence in distribution, probability, and almost surely (often abbreviated "a.s."), respectively. For random vectors or matrices, we use the same notation to denote such convergence, entry by entry. Let Π_N denote the set of permutations of $\{1, \ldots, N\}$. Let π denote a generic element of Π_N, which is a mapping from $\{1, \ldots, N\}$ to itself. Let $\text{Unif}(\Pi_N)$ denote the uniform distribution over Π_N. Random variable B stochastically dominates A, written $A \leq_{\text{st}} B$, if their cumulative distribution functions $F_A(x)$ and $F_B(x)$ satisfy $F_A(x) \geq F_B(x)$ for all x. Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed (i.i.d.) $\mathcal{N}(0, 1)$ random variables.

Chapter 2

Framework for Randomization Tests

2.1 Completely Randomized Experiments

We adhere to the potential outcomes framework of [72] and [87]. Let $Y_i(j)$ be the response of unit i if it receives treatment j, where $i = 1, \ldots, N$ and $j = 1, \ldots, J$. Vectorize $Y_i = (Y_i(1), \ldots, Y_i(J))^\top$. The means of the potential outcomes are $\bar{Y}(j) = \sum_{i=1}^N Y_i(j)/N$, vectorized as $\bar{Y} = (\bar{Y}(1), \ldots, \bar{Y}(J))^\top$. The covariance between the potential outcomes is $S(j,k) = \sum_{i=1}^N \{Y_i(j) - \bar{Y}(j)\}\{Y_i(k) - \bar{Y}(k)\}/(N-1)$, which is a variance if $j = k$. The covariance matrix S has the (j,k)-th entry $S(j,k)$.

Let $W_i \in \{1, \ldots, J\}$ represent the treatment that unit i actually receives, and define the indicator $W_i(j) = 1(W_i = j)$. The $W_1, \ldots, W_N$ are generated according to a completely randomized experiment (CRE). The experimenter picks $N_1, \ldots, N_J \geq 2$ that sum to N, and assigns treatments randomly so that any realization satisfies $\sum_{i=1}^N W_i(j) = N_j$ for $j = 1, \ldots, J$, and has probability $\prod_{j=1}^J N_j!/N!$.

Unit i's observed outcome is $Y_i^{\text{obs}} = Y_i(W_i) = \sum_{j=1}^J W_i(j) Y_i(j)$. So the observed means are $\hat{\bar{Y}}(j) = \sum_{i=1}^N W_i(j) Y_i^{\text{obs}}/N_j$, vectorized as $\hat{\bar{Y}} = (\hat{\bar{Y}}(1), \ldots, \hat{\bar{Y}}(J))^\top$. The observed variances are $\hat{S}(j,j) = \sum_{i=1}^N W_i(j)\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}^2/(N_j - 1)$, which is the sample analog of $S(j,j)$. Because $Y_i(j)$ and $Y_i(k)$ are not jointly observable, there is no sample analog for $S(j,k)$. In general, we cannot estimate $S(j,k)$ consistently for $j \neq k$. For regularity, we assume $S(j,j) > 0$ and $\hat{S}(j,j) > 0$ for all $W = (W_1, \ldots, W_N)^\top$.

2.2 Fisher Randomization Tests

The *Fisher Randomization Test* (FRT) was formulated by [30] to analyze experimental data. Several variations of it exist [77, 42, 8, 79]. We adopt that of [86]. It arises from the potential outcomes described in the previous section.

[85] called the potential outcome matrix $\{Y_i(j) : i = 1, \ldots, N, j = 1, \ldots, J\}$ the Science Table. He termed a null hypothesis *sharp* if it, along with the observed data, can determine all the missing items in the Science Table. A test statistic is a function of the observed data and the null hypothesis. Under a sharp null hypothesis, any test statistic has a known

distribution. In particular, we can cycle through the possible values of W, and for each obtain the corresponding realization of observed data, and then compute the value of the test statistic. In this manner, the test statistic's distribution becomes accessible, as does a p-value. FRTs are therefore finite-sample exact for testing sharp null hypotheses, no matter the test statistic or data generating process for the potential outcomes [83, 46]. In essence, randomization is fundamental for statistical inference. It justifies the FRT, and guarantees the validity of the resulting p-value.

Practitioners typically brand sharp null hypotheses as too restrictive. In a general factorial experiment, our mission is to test

$$H_{0\mathrm{N}}(C, x) : C\bar{Y} = x, \tag{2.1}$$

where $x \in \mathbb{R}^m$ and $C \in \mathbb{R}^{m \times J}$ is a contrast matrix of full row rank m, i.e., $C1_J = 0_m$. We pay particular attention to hypotheses where $x = 0_m$, but study general x for completeness. A *weak* hypothesis is any that is not sharp by the definition of [85]. The hypothesis (2.1) is therefore weak. It is also referred to as an average/Neyman null hypothesis. It only confines the averages of the potential outcomes. On the other hand, a sharp/strong/Fisher null hypothesis confines the individual potential outcomes.

Notwithstanding that the FRT is designed for sharp null hypotheses, we ask whether it can test (2.1) also. The FRT mandates that all potential outcomes be filled out. We do so aided by an artificial sharp null hypothesis. A sensible one is

$$H_{0\mathrm{F}}(C, x, \tilde{C}, \tilde{x}) : \begin{pmatrix} C \\ \tilde{C} \end{pmatrix} Y_i = \begin{pmatrix} x \\ \tilde{x} \end{pmatrix} \text{ for } i = 1, \ldots, N, \tag{2.2}$$

where the matrix $(C^\top, \tilde{C}^\top, 1_J)$ is invertible. Given C and 1_J, we can construct $\tilde{C}$ from Gram-Schmidt orthogonalization. When $m < J - 1$, we are also to pick a value $\tilde{x} \in \mathbb{R}^{J-m-1}$. In the case with $x = 0_m$, we can go with $\tilde{x} = 0_{J-m-1}$ to get the classical sharp null hypothesis of no individual effects whatsoever. Intuitively, the piece $CY_i = x$ of (2.2) is "of" the weak null hypothesis (2.1), and the piece $\tilde{C}Y_i = \tilde{x}$ is "beyond" it. The hypothesis (2.2) induces two key features. The first is the weak null hypothesis (2.1). The second is strict additivity, i.e., $Y_i(j) - Y_i(k)$ does not depend on the unit i, for $j, k = 1, \ldots, J$.

With the sharp null hypothesis (2.2) and some test statistic T that ideally can capture possible deviation from (2.2), the FRT proceeds as follows.

FRT-1. Calculate T from $\{W_i, Y_i^{\mathrm{obs}} : i = 1, \ldots, N\}$.

FRT-2. Impute potential outcomes:

$$Y_i^* = \begin{pmatrix} Y_i^*(1) \\ \vdots \\ Y_i^*(J) \end{pmatrix} = z + (Y_i^{\mathrm{obs}} - z_{W_i})1_J, \text{ where } z = \begin{pmatrix} z_1 \\ \vdots \\ z_J \end{pmatrix} = \begin{pmatrix} C \\ \tilde{C} \\ 1_J^\top \end{pmatrix}^{-1} \begin{pmatrix} x \\ \tilde{x} \\ 0 \end{pmatrix},$$

or, equivalently, $Y_i^*(j) = Y_i^{\mathrm{obs}} + z_j - z_{W_i}$ for $j = 1, \ldots, J$.

FRT-3. For a permutation $\pi \in \Pi_N$, compute $Y_{\pi,i}^{\text{obs}} = \sum_{j=1}^{J} W_{\pi(i)}(j) Y_i^*(j)$ and calculate T_π from $\{W_{\pi(i)}, Y_{\pi,i}^{\text{obs}} : i = 1, \ldots, N\}$ the same way T was calculated.

FRT-4. The p-value is $(N!)^{-1} \sum_{\pi \in \Pi_N} 1(T_\pi \geq T)$.

As a sanity check, the imputed potential outcomes in FRT-2 satisfy (2.2) and $Y_i^*(W_i) = Y_i^{\text{obs}}$ for all i.

Given the Science Table, every realization of treatment assignment W produces data $\{W_i, Y_i^{\text{obs}} : i = 1, \ldots, N\}$. Henceforth, we call the values of T that can possibly emerge from these data the *randomization distribution* of T. Conditioning on the original data $\{W_i, Y_i^{\text{obs}} : i = 1, \ldots, N\}$, we can fill out missing potential outcomes with FRT-2. We call the set of values $\{T_\pi : \pi \in \Pi_N\}$ defined in FRT-3 the *permutation distribution* of T. Since this distribution depends on the original data, whose randomness comes solely from W, we denote this distribution with $T_\pi | W$.

If the treatment truly does not affect any unit whatsoever, then the FRT just described reduces to the classical permutation test. In these circumstances, the FRT and permutation test are numerically identical. There is an isomorphism between the two in this sense. In general, the FRT admits a broader class of null hypotheses and experimental designs than the permutation test.

Step FRT-4 conveys that the FRT p-value is a right-tail probability. A larger value of T embodies a larger deviation from the null hypothesis. Even if $N!$ is too large for a manageable exact computation of the p-value, we are able to fall back on random iid draws from Π_N to approximate the p-value in FRT-4 subject to Monte Carlo error. We are thus always at liberty to sample randomly from the permutation distribution.

For any test statistic T, the p-value in FRT-4 is valid under (2.2). Our central goal is to investigate whether the FRT can still control type I error for testing $H_{0\text{N}}(C, x)$. Roughly speaking, this turns out to be affirmative asymptotically with an appropriate test statistic T. Before continuing, let us be specific that the FRT with T successfully controls type I error at level α if $\mathbb{P}\left\{(N!)^{-1} \sum_{\pi \in \Pi_N} 1(T_\pi \geq T) \leq \alpha\right\} \leq \alpha$. When the probability could be smaller than α, we might say the test is *conservative* for added emphasis. However, a conservative test still successfully controls type I error.

2.3 Asymptotics for Finite Population Inference

We have contended that the exact randomization distribution of T depends on unknown potential outcomes under $H_{0\text{N}}(C, x)$ in general. Finite-sample theory in this case is not feasible. Instead, we embrace an asymptotic theory. Imagine a sequence of finite populations of potential outcomes. For each $N \geq 2J$, we fix in advance $N_1, \ldots, N_J \geq 2$. Independently across N, we generate W according to a CRE, from which we get Y_i^{obs} and calculate a test statistic. We denote a sequence indexed by N with $N \to \infty$ by $(\cdot)$ or $(\cdot)_{N \geq 2J}$. Technically, we should index finite population quantities by N, and also index observed quantities by $N_1, \ldots, N_J$. For cleaner notation, and following the precedent of earlier authors, we

drop these extra subscripts, unless to emphasize the dependence on N. We now state our assumptions on the sequence of potential outcomes.

Assumption A. The sequence (N_j/N) converges to $p_j \in (0,1)$ for all $j = 1, \ldots, J$. The sequences $(\bar{Y}_N)$ and (S_N) converge to $\bar{Y}_\infty < \infty$ and S_∞, where S_∞ has finite entries and positive main diagonal entries. Further, $\lim_{N\to\infty} \max_{j=1,\ldots,J} \max_{i=1,\ldots,N} \left\{Y_i(j) - \bar{Y}(j)\right\}^2/N = 0$.

Assumption B. Same as Assumption A with the last equation replaced by: there exists an $L < \infty$ such that $\sum_{i=1}^{N} \{Y_i(j) - \bar{Y}(j)\}^4/N \leq L$ for all $j = 1, \ldots, J$ and $N \geq 2J$.

Proposition 1. *Assumption B implies Assumption A.*

The design of experiments often guarantees the existence of $p_j \in (0,1)$ because all treatment groups have comparable sizes in realistic cases. We can weaken the existence of $\bar{Y}_\infty$ and S_∞ by standardizing the potential outcomes. Just as we drop N, we might drop subscripts ∞. For instance, S can mean either the finite population covariance matrix or its limiting value, which will be clear from context. Intuitively, Assumption A requires more than two moments, and Assumption B requires four moments. Assumption B is thus stronger than Assumption A. Below are our principal asymptotic tools, which are consequences of [58].

Proposition 2. *Under Assumption A,* $\hat{\bar{Y}} - \bar{Y} \xrightarrow{P} 0_J$, *and* $\hat{S}(j,j) \xrightarrow{P} S(j,j)$ for $j = 1, \ldots, J$.

Proposition 3. *Under Assumption A,* $N^{1/2}(\hat{\bar{Y}} - \bar{Y}) \xrightarrow{\mathrm{d}} \mathcal{N}(0_J, V)$, *where*

$$V = \lim_{N\to\infty} N \cdot \mathrm{Cov}(\hat{\bar{Y}}) = \lim_{N\to\infty} \begin{pmatrix} \frac{N-N_1}{N_1} S(1,1) & -S(1,2) & \cdots & -S(1,J) \\ -S(2,1) & \frac{N-N_2}{N_2} S(2,2) & \cdots & -S(2,J) \\ \vdots & \vdots & \ddots & \vdots \\ -S(J,1) & -S(J,2) & \cdots & \frac{N-N_J}{N_J} S(J,J) \end{pmatrix}. \tag{2.3}$$

The limiting distribution in Proposition 3 depends on unknown quantities. We need to estimate $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$. This covariance, however, depends on $S(j,k)$ $(j \neq k)$, which do not have unbiased estimators in general. Estimating the main diagonal is the best we can hope to do:

$$\hat{D} = N \cdot \mathrm{diag}\left\{\hat{S}(1,1)/N_1, \cdots, \hat{S}(J,J)/N_J\right\} \succ 0.$$

Proposition 2 implies

$$\hat{D} \xrightarrow{P} D = \mathrm{diag}\left\{S(1,1)/p_1, \cdots, S(J,J)/p_J\right\} \succ 0. \tag{2.4}$$

Therefore, $V = D - S \preceq D$. $\hat{D}$ is an asymptotically conservative estimator for $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$ in the sense that $\lim_{N\to\infty} N \cdot \mathrm{Cov}(\hat{\bar{Y}}) \preceq \mathrm{plim}_{N\to\infty} \hat{D}$. We will encounter this notion repeatedly.

Chapter 3

Test Statistics

3.1 Proper Test Statistics

We return to our main endeavor: whether the FRT with a test statistic T can control type I error when testing $H_{0\mathrm{N}}(C, x)$. The next proposition demarcates precisely what kind of T can accomplish this goal.

Proposition 4. *Consider testing $H_{0\mathrm{N}}(C, x)$. The FRT with test statistic T controls type I error at any level if and only if, under $H_{0\mathrm{N}}(C, x)$, we have $T \leq_{\mathrm{st}} T_\pi|W$ a.s., that is, if and only if the randomization distribution of T is stochastically dominated by its permutation distribution.*

To test $H_{0\mathrm{N}}(C, x)$, we use a test statistic T, but look upon its permutation distribution $T_\pi|W$ as the reference null distribution. The p-value in FRT-4 is the probability that $T_\pi|W$ is at least the observed value of T. If $T \leq_{\mathrm{st}} T_\pi|W$, then any quantile of the asymptotic distribution of $T_\pi|W$ is at least that of T. Consequently, we have asymptotically conservative tests at any level.

It is too unwieldy to ensure a meaningful test statistic satisfies the criterion of Proposition 4. For a candidate statistic T, we instead settle for ascertaining whether its permutation distribution stochastically dominates its randomization distribution asymptotically under $H_{0\mathrm{N}}(C, x)$ for almost all sequences of W. Henceforth, we call T *proper* if so.

3.2 Studentized statistic

We advocate using the following studentized statistic in the FRT:

$$X^2 = N(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1}(C\hat{\bar{Y}} - x). \tag{3.1}$$

It is a Wald-type statistic with a conservative covariance estimator $C\hat{D}C^\top$ for $N^{1/2}(C\hat{\bar{Y}} - x)$.

Studentized statistics have appeared alongside permutation tests in an independent samples setting. [79] was aware of the problem of statistics that lacked studentization in two-sample tests. For [48], studentization was an avenue in the Behrens–Fisher problem to control

the type I error. [19] studied the same phenomenon when the parameter of interest could be more general than the mean. [76] and [54] embraced an equivalent studentized statistic in general factorial experiments with independent samples. In the aforementioned settings, studentization works because the test statistic is asymptotically pivotal.

As for us, X^2 is itself not asymptotically pivotal. Rather, it is stochastically dominated by a pivotal distribution. This is a key reason it is exactly the statistic we seek based on Proposition 4. We now formally state our main result that X^2 is proper.

Theorem 1. *If Assumption A holds, then under* $H_{0\mathrm{N}}(C, x)$, $X^2 \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2$, *where each* $a_j \in [0, 1]$. *If Assumption B holds, and* $\pi \sim \mathrm{Unif}(\Pi_N)$, *then* $X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_m^2$ *a.s.*

Immediate from this theorem is that the FRT using X^2 controls the asymptotic type I error under $H_{0\mathrm{N}}(C, x)$. This test also retains finite sample exactness under any sharp null hypothesis. As a result, it is robust for inference on two classes of null hypotheses.

Asymptotically, under $H_{0\mathrm{N}}(C, x)$, neither the randomization nor permutation distribution of X^2 depends on $\tilde{C}$ or $\tilde{x}$, so the choice of $\tilde{x}$ does not matter. The permutation distribution also does not depend on $H_{0\mathrm{N}}(C, x)$. A violation of $H_{0\mathrm{N}}(C, x)$ is likely to inflate the value of X^2 but not the values of $X_\pi^2 | W$. An appealing consequence of this fact is that the FRT using X^2 has power.

Echoing [19] and [76], one purpose of studentization for us is to control type I error. Yet, for us, the FRT using X^2 is asymptotically conservative, while the corresponding test in an independent samples setting is asymptotically exact. This stems from our potential outcomes framework: $\{\hat{\tilde{Y}}(1), \ldots, \hat{\tilde{Y}}(J)\}$ do not have vanishing correlations, even asymptotically.

Theorem 1 inspires another asymptotically conservative test besides the FRT. We can reject $H_{0\mathrm{N}}(C, x)$ if the observed value of X^2 exceeds the $1 - \alpha$ quantile of χ_m^2. We call this alternative to the FRT the χ^2 approximation. This is computationally efficient without Monte Carlo. The FRT has an additional property. It is concurrently finite-sample exact for any sharp null hypothesis. Our simulations and practical data examples compare these two classes of tests empirically.

3.3 Box-Type Statistic

We now steer toward an alternative statistic, one found in [15]:

$$B = N \hat{\tilde{Y}}^\top M \hat{\tilde{Y}} / \operatorname{tr}(M \hat{D}), \tag{3.2}$$

where $M = C^\top (C C^\top)^{-1} C$ is the projection matrix onto the row space of C. Because we will deem it as not proper in our context, we can restrict the discussion to $x = 0_m$.

Under independent sampling, [15] approximated the asymptotic behavior of B by an F distribution through ideas from [12], and called it a Box-type statistic. Their simulations found it to enjoy superior empirical small sample properties under their framework.

For our problem, the next result states the behavior of B. Recall V in (2.3) and define $P = \mathrm{diag}(p_1, \ldots, p_J)$.

Theorem 2. *If Assumption A holds, then under* $H_{0N}(C, 0_m)$, $B \xrightarrow{d} \sum_{j=1}^{m} \lambda_j(MV)\xi_j^2 / \operatorname{tr}(MD)$. *If Assumption B holds and* $\pi \sim \mathrm{Unif}(\Pi_N)$, *then* $B_\pi | W \xrightarrow{d} \sum_{j=1}^{m} \lambda_j(MP^{-1})\xi_j^2 / \operatorname{tr}(MP^{-1})$ *a.s.*

The asymptotic mean of B is $\sum_{j=1}^{m} \lambda_j(MV)/\operatorname{tr}(MD) \leq 1$ because $V \preceq D$, and the asymptotic mean of $B_\pi | W$ is $\sum_{j=1}^{m} \lambda_j(MP^{-1})/\operatorname{tr}(MP^{-1}) = 1$. Therefore, the former mean does not exceed the latter. This is necessary but not sufficient for the stochastic dominance criterion of Proposition 4, which does not hold. Hence, the FRT with the Box-type statistic cannot control type I error in general, even asymptotically. This is the subject of a later simulation.

There are two situations where B is proper: equal variances, and testing a one-dimensional hypothesis.

Corollary 1. *Under Assumption B, if* $S(1,1) = \cdots = S(J,J)$, *then* B *meets the criterion of Proposition 4 asymptotically. If* C *is a row vector, then* $B = X^2$.

3.4 Statistics from Ordinary Least Squares

Ordinary least squares (OLS) tools are widespread in the analysis of experimental data [e.g., 68]. To fit J-treatment randomized experiments into the linear models framework, the design matrix $\mathcal{X} = \operatorname{diag}\{1_{N_1}, \ldots, 1_{N_J}\}$ is block diagonal. The response vector consists of the corresponding observed outcomes from treatment groups $1, \ldots, J$. The OLS coefficients are the entries of $\hat{\hat{Y}}$, which has estimated covariance matrix $\hat{\sigma}^2(\mathcal{X}^\top\mathcal{X})^{-1}$, where $\hat{\sigma}^2 = (N - J)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{J} W_i(j)\{Y_i^{\mathrm{obs}} - \hat{\hat{Y}}(j)\}^2$ is the mean residual sum of squares. Based on these, the classical F statistic is

$$F = (C\hat{\hat{Y}})^\top \{\hat{\sigma}^2 C(\mathcal{X}^\top\mathcal{X})^{-1}C^\top\}^{-1} C\hat{\hat{Y}}/m. \tag{3.3}$$

We do not impose the usual assumptions of linear regression, but just want the test statistic F.

We first record a peculiar situation where F is identical to the Box-type statistic B. This result will be valuable for our simulations and practical data examples.

Proposition 5. $B = F$ *if* $N_1 = \cdots = N_J$ *and* $M = C^\top(CC^\top)^{-1}C$ *has the same entries along its main diagonal.*

Except for the scaling by m and the presence of $\hat{\sigma}^2$ in place of each $\hat{S}(j,j)$, F is identical to X^2. This pooled variance estimate $\hat{\sigma}^2$ is problematic for the F statistic problematic, as we formalize next.

Theorem 3. *If Assumption A holds, then under* $H_{0N}(C, 0_m)$,

$$m \cdot F \xrightarrow{d} \sum_{j=1}^{m} \lambda_j\big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big)\xi_j^2,$$

where $\bar{S} = \sum_{j=1}^{J} p_j S(j,j)$. *If Assumption B holds and* $\pi \sim \mathrm{Unif}(\Pi_N)$, *then* $m \cdot F_\pi | W \xrightarrow{d} \chi^2_m$ *a.s.*

The classical linear model assumes a constant treatment effect for all units [52]. This necessitates equal variances under all treatment levels. Yet, such homoscedasticity is not built into the potential outcomes framework. The assumptions underlying the F statistic are not compatible with the potential outcomes framework in general. If the potential outcomes do have equal variance, then it is not surprising that F is proper.

Corollary 2. *Under Assumption B, if $S(1,1) = \cdots = S(J,J)$, then F meets the criterion of Proposition 4 asymptotically.*

Huber–White covariance estimation for the OLS coefficients [44, 94] is frequently quoted as a fix to the classical F statistic. Econometricians in particular are inclined to such an estimate of the covariance when the linear model is possibly misspecified or the error terms are heteroscedastic. Define the residual $\hat{\epsilon}_i = Y_i^{\text{obs}} - \hat{\bar{Y}}(W_i)$. The Huber–White estimator for $N \cdot \text{Cov}(\hat{\bar{Y}})$ is

$$\begin{aligned}\hat{D}_{\text{HW}} =& N(\mathcal{X}^\top\mathcal{X})^{-1}\mathcal{X}^\top \operatorname{diag}\left\{\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_N^2\right\}\mathcal{X}(\mathcal{X}^\top\mathcal{X})^{-1} \\ =& N \cdot \operatorname{diag}\left\{\frac{N_1 - 1}{N_1^2}\hat{S}(1,1), \ldots, \frac{N_J - 1}{N_J^2}\hat{S}(J,J)\right\}.\end{aligned}$$

If we replace $\hat{\sigma}^2(\mathcal{X}^\top\mathcal{X})^{-1}$ by $\hat{D}_{\text{HW}}$ in (3.3) and dismiss the scaling by m, we get

$$X^2_{\text{HW}} = N(C\hat{\bar{Y}})^\top(C\hat{D}_{\text{HW}}C^\top)^{-1}C\hat{\bar{Y}}.$$

$\hat{D}_{\text{HW}}$ is nearly identical to $\hat{D}$ if $N_j \approx N_j - 1$ for $j = 1, \ldots, J$. Therefore, X^2_{HW} is asymptotically akin to X^2. By this, the Huber–White covariance estimator successfully repairs the F statistic.

Chapter 4

Special Cases

Section 3 devises a strategy for testing weak null hypotheses in general experiments. The results are directly applicable to many worthwhile settings.

4.1 One-Way Analysis of Variance

In the one-way analysis of variance (ANOVA), the goal is to test $H_{0\mathrm{N}} : \bar{Y}(1) = \cdots = \bar{Y}(J)$. It is a special case of the null hypothesis (2.1) with any contrast matrix $C \in \mathbb{R}^{(J-1)\times J}$ and $x = 0_{J-1}$, for instance $C = (1_{J-1}, -I_{J-1})$. In this case, we can impute potential outcomes in FRT-2 as $Y_i^*(j) = Y_i^{\mathrm{obs}}$ for $i = 1, \ldots, N$ and $j = 1, \ldots, J$ under $H_{0\mathrm{F}} : Y_i(1) = \cdots = Y_i(J)$, for $i = 1, \ldots, N$. To test $H_{0\mathrm{F}}$, [29] crafted the statistic

$$F = \frac{\sum_{j=1}^{J} N_j \{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\mathrm{obs}}\}^2/(J-1)}{\sum_{j=1}^{J}(N_j - 1)\hat{S}(j,j)/(N-J)}, \text{ where } \bar{Y}_\bullet^{\mathrm{obs}} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{\mathrm{obs}}. \tag{4.1}$$

He argued that $F_{J-1,N-J}$ approximates the randomization distribution of F. [24] attested that (4.1) is not proper but

$$X^2 = \sum_{j=1}^{J} \frac{N_j}{\hat{S}(j,j)} \{\hat{\bar{Y}}(j) - \bar{Y}_S^{\mathrm{obs}}\}^2, \text{ where } \bar{Y}_S^{\mathrm{obs}} = \frac{\sum_{j=1}^{J} N_j \hat{\bar{Y}}(j)/\hat{S}(j,j)}{\sum_{j=1}^{J} N_j/\hat{S}(j,j)} \tag{4.2}$$

is for testing $H_{0\mathrm{N}}$ with the FRT. See [91] for a related discussion.

It is immediate from the next proposition that our framework encompasses these results as special cases.

Proposition 6. *In the one-way ANOVA, the* X^2 *in* (3.1) *and* (4.2) *coincide, as do the* F *in* (3.3) *and* (4.1).

4.2 Treatment-Control Experiments

In the treatment-control setting, $J = 2$, and unit i either receives the treatment (then $Y_i^{\mathrm{obs}} = Y_i(1)$) or control (then $Y_i^{\mathrm{obs}} = Y_i(2)$). A parameter of interest is the average treatment

effect $\tau = \bar{Y}(1) - \bar{Y}(2)$. The weak null hypothesis is $H_{0\mathrm{N}}(C, 0) : \tau = 0$. This matches (2.1), where $C = (1, -1)$ is a row vector. Thus, treatment-control is a special case of the one-way layout. A popular statistic is $|\hat{\tau}|$, where $\hat{\tau} = \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)$ is the sample difference-in-means of outcomes. However, [24] showed that $|\hat{\tau}|$ is not proper for testing $H_{0\mathrm{N}}$ in general.

Corollary 3. *Under Assumption B, in the treatment-control setting, for almost all sequences of W, $B = X^2$ can asymptotically control type I error, but F and $|\hat{\tau}|$ cannot, unless $N_1 = N_2$ or $S(1, 1) = S(2, 2)$.*

From Corollary 1, the Box-type statistic B equals the studentized statistic X^2 in the treatment-control setting. Both are proper. The statistic $|\hat{\tau}|$ is not only not proper, but it also has other "paradoxical" shortcomings [23]; see also the comment of [61]. Corollary 3 declares that a balanced design can salvage the F and $|\hat{\tau}|$ statistics, even without homoscedasticity. Perhaps counter to intuition, this does not stay true when $J > 2$, as our simulations later demonstrate.

4.3 Trend Tests

Our perspective has been on type I error under null hypotheses without specifying alternative hypotheses. In experiments for dose-response relationships, we have ordered treatment $1 \leq \cdots \leq J$ and often specify the null and alternative hypotheses as $H_{0\mathrm{N}}$ and $H_{1\mathrm{N}} : \bar{Y}(1) \leq \cdots \leq \bar{Y}(J)$ with at least one strict inequality. We can still carry forward the results in Section 4.1 on ANOVA. Power might shrink for the test if we do not account for the ordering of the dose-response relationship. Motivated by [3] and [75], we first choose doses $(a_1, \ldots, a_J)$ for treatment levels $(1, \ldots, J)$. Then the test statistic

$$r = \sum_{j=1}^{J} a_j \{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\mathrm{obs}}\} = C\hat{\bar{Y}},$$

is plausible, where $C = (a_1 - a_+ N_1/N, \ldots, a_J - a_+ N_J/N) \in \mathbb{R}^{1\times J}$ is a contrast vector, and $a_+ = \sum_{j=1}^{J} a_j$. In effect, we are testing $H_{0\mathrm{N}}(C, 0) : C\bar{Y} = 0$. Previous theory suggests that r is not proper but the studentized statistic is:

$$t = C\hat{\bar{Y}}/(C\hat{D}C^\top/N)^{1/2}.$$

Note that under $H_{0\mathrm{N}}$, we impute all missing potential outcomes as Y_i^{obs} for each unit i, albeit we fix a particular contrast vector C to construct the studentized statistic. Moreover, in this case, we conduct a one-sided test, rejecting $H_{0\mathrm{N}}$ if t is larger than the $1 - \alpha$ quantile of its permutation distribution.

4.4 Binary Outcomes

The theory for X^2 statistics does not insist that the outcome be of a particular type as long as the regularity conditions hold. In particular, it applies directly to binary outcomes.

However, binary outcomes have a special feature that $S(j,j) = N\bar{Y}(j)\{1-\bar{Y}(j)\}/(N-1)$, i.e., the mean $\bar{Y}(j)$ determines the variance $S(j,j)$. Therefore, under the null hypothesis $H_{0\text{N}} : \bar{Y}(1) = \cdots = \bar{Y}(J)$, the variances are all the same too: $S(1,1) = \cdots = S(J,J)$. With binary outcomes, the difference-in-means statistic $|\hat{\tau}|$ for $J = 2$ in Section 4.2, the F statistic for general J in Section 4.1, and the r statistic in Section 4.3 are all proper, for testing $H_{0\text{N}}$. As pointed out by [23], for this weak null hypothesis, we do not need studentization to guarantee correct asymptotic type I error. However, this does not hold for general weak null hypotheses $H_{0\text{N}}(C, x)$ of binary potential outcomes because $C\bar{Y} = x$ does not imply they have equal variances. In general, we always recommend using X^2.

4.5 2^K Factorial Designs

2^K factorial designs seek to analyze K binary treatment factors simultaneously. In total, we have $J = 2^K$ possible treatment combinations. [22] tied these designs together with the potential outcomes framework. We summarize this setup. To do so, it is helpful to introduce the model matrix $G \in \{\pm 1\}^{(J-1)\times J}$. Let $*$ denote the component-wise product. [63] constructed the rows of G, which we call $g_1^\top, \ldots, g_{J-1}^\top$, as follows:

- for $j = 1, \ldots, K$, let $g_j^\top$ be $-1_{2^{K-j}}^\top, 1_{2^{K-j}}^\top$ repeated 2^{j-1} times;
- the next $\binom{K}{2}$ values of g_j's are $g_{k(1)} * g_{k(2)}$ where $k(1) \neq k(2) \in \{1, \ldots, K\}$;
- the next $\binom{K}{3}$ are component-wise products of triplets of distinct $g_1, \ldots, g_K$, etc;
- the bottom row is $g_{J-1} = g_1 * \cdots * g_K$.

The matrix G has rows orthogonal to each other and to 1_J, i.e., $GG^\top = J \cdot I_{J-1}$ and $G1_J = 0_{J-1}$. Let $\tilde{G} \in \{\pm 1\}^{K\times J}$ be the first K rows of G. Call its columns $z_1, \ldots, z_J$, which are the possible treatment combinations. The following example elucidates the setup.

Example 1. When $K = 2$, we have

$$G = \begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} g_1^\top \\ g_2^\top \\ g_3^\top \end{pmatrix} = \begin{pmatrix} \tilde{G} \\ g_3^\top \end{pmatrix} = \begin{pmatrix} z_1 & z_2 & z_3 & z_4 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

The four possible treatment combinations are $z_1 = (-1,-1)^\top$, $z_2 = (-1,1)^\top$, $z_3 = (1,-1)^\top$, and $z_4 = (1,1)^\top$. We read these off from the first two rows of G. □

The rows of G define factorial effects. Namely, $g_1, \ldots, g_K$ correspond to main effects, $g_{K+1}, \ldots, g_{K+\binom{K}{2}}$ correspond to two-way interactions, etc, and g_{J-1} corresponds to the K-way interaction. Let $Y_i(j) = Y_i(z_j)$ be the response of unit i if it receives the treatment combination z_j. Then we can transfer our previous notation to 2^K factorial designs. The general factorial effect for unit i indexed by g_j is $\tau_{ij} = 2g_j^\top Y_i/J$, and the corresponding average factorial effect is $\bar{\tau}_{\bullet j} \sum_{i=1}^N \tau_{ij}/N = 2g_j^\top \bar{Y}/J$. Vectorize these quantities: $\tau_i = (\tau_{i1}, \ldots, \tau_{i,J-1})^\top = 2GY_i/J$ and $\tau = (\bar{\tau}_{\bullet 1}, \ldots, \bar{\tau}_{\bullet J-1})^\top = 2G\bar{Y}/J$.

We may perform inference on τ or any subset of its entries. Let $A = \{a(1), \ldots, a(m)\} \subseteq \{1, \ldots, J-1\}$ be the target subset, and let $C \in \{\pm 1\}^{m \times J}$ have rows $g_{a(1)}^\top, \ldots, g_{a(m)}^\top$. Then $\tau_A = (\bar{\tau}_{\bullet a(1)}, \ldots, \bar{\tau}_{\bullet a(m)})^\top = 2C\bar{Y}/J$. Testing whether $\tau_A = 2x/J$ is equivalent to testing $H_{0\mathrm{N}}(C, x)$. The FRT with X^2 is proper. The factorial design stimulates a natural choice of $\tilde{C}$ for the imputation step FRT-2. We let $g_j^\top$ be a row of $\tilde{C}$ whenever $j \notin A$.

[63] discussed both randomization-based and regression-based inferences for 2^K factorial designs. He fixated on point estimation and proposed using the Huber–White covariance estimator. We have likewise highlighted that it is imperative to use the Huber–White covariance estimator and the F statistic together in the FRT.

4.6 Hodges–Lehmann Estimation

Up to this stage, our developments have been on hypothesis testing. Drawing upon the duality between testing and estimation, our previous results shed light on the estimation of $C\bar{Y}$. This strategy is sometimes referred to as Hodges–Lehmann estimation [41, 83]. For a fixed x, we can by means of the FRT obtain a p-value for the null hypothesis $H_{0\mathrm{N}}(C, x)$. Let us denote this p-value by $p(x)$ to delineate its dependence on x.

The Hodges–Lehmann point estimator $\hat{\tau}_{\mathrm{HL}}$ for $C\bar{Y}$ is the $x \in \mathbb{R}^m$ that results in the least significant p-value for testing $H_{0\mathrm{N}}(C, x)$. In symbols, $\hat{\tau}_{\mathrm{HL}} \in \operatorname{argmax}_{x \in \mathbb{R}^m} p(x)$. Note that $x = C\hat{\bar{Y}}$ implies $X^2 = 0$, which in turn implies $p(x) = 1$. Thus $\hat{\tau}_{\mathrm{HL}} = C\hat{\bar{Y}}$, the usual unbiased estimator. Because X^2 is proper, the duality between hypothesis testing and confidence sets assures us that

Corollary 4. *For $\alpha \in (0, 1)$ and almost all sequences of W, an asymptotically conservative $(1-\alpha)$ confidence set for $C\bar{Y}$ is*

$$\mathrm{CR}_\alpha = \big\{x \in \mathbb{R}^m : p(x) > \alpha\big\},$$

in the sense that $\lim_{N\to\infty} \mathbb{P}\{C\bar{Y} \in \mathrm{CR}_\alpha\} \geq 1 - \alpha$.

Determining CR_α can be computationally intensive, so it is expedient to have the asymptotic approximation

$$\mathrm{CR}_\alpha \approx \left\{x : N(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1}(C\hat{\bar{Y}} - x) \leq \chi^2_{m,\alpha}\right\}, \tag{4.3}$$

where $\chi^2_{m,\alpha}$ is the $1-\alpha$ quantile of χ^2_m. Because the X^2 statistic is a quadratic form, CR_α is an ellipsoid centered at $C\hat{\bar{Y}}$. The set CR_α can serve either directly as a $1-\alpha$ approximate confidence set or as an initial guess in searching for the exact confidence region by inverting FRTs. We undertake this later by a simulation.

4.7 Testing Inequalities

FRTs can also handle hypotheses of inequalities:

$$\tilde{H}_{0\mathrm{N}}(C, x) : C\bar{Y} \geq x. \tag{4.4}$$

We commence with the case where $C \in \mathbb{R}^{1\times J}$ is a row vector with $C1_J = 0$, and $x \in \mathbb{R}$.

Example 2. In the two-sample problem with $J = 2$, we can test $\bar{Y}(2) - \bar{Y}(1) \geq 0$: whether treatment level 1 results in smaller outcomes than treatment level 2 on average. In this case, $C = (-1, 1)$ and $x = 0$. □

Example 3. In a gold standard design for three arms, let level 1 be the placebo control, level 2 be the active control, and level 3 be the experimental treatment. Suppose that smaller outcomes are more desirable, and we know that $\bar{Y}(2) > \bar{Y}(1)$ from previous studies. Given $\Delta > 0$, the goal is to test the hypothesis $\bar{Y}(1) - \bar{Y}(3) \leq \Delta\{\bar{Y}(1) - \bar{Y}(2)\}$. When $\Delta > 1$, this is a superiority test, and when $\Delta \in (0, 1)$, this is a non-inferiority test [70]. This null hypothesis is equivalent to $\tilde{H}_{0\mathrm{N}}(C, 0) : (\Delta - 1)\bar{Y}(1) - \Delta\bar{Y}(2) + \bar{Y}(3) \geq 0$ with $C = (\Delta - 1, -\Delta, 1)$. □

To impute the missing potential outcomes, we pretend that the null hypothesis is $H_{0\mathrm{N}}(C, x)$ and utilize (2) as we did before. The statistic X^2 is not suitable here because it is intended for two-sided tests. For instance, X^2 can be large, even under $\tilde{H}_{0\mathrm{N}}(C, x)$. Instead we use a truncated statistic $t_+ = \max(t, 0)$ where

$$t = N^{1/2}(x - C\hat{\bar{Y}})/(C\hat{D}C^\top)^{1/2}.$$

The FRT with t also works for p-values at most 0.5. [70] used the special case of t in the setting of Example 3. We choose t_+ so that Proposition 4 directly covers our situation. We summarize the results below.

Corollary 5. *Consider testing $\tilde{H}_{0\mathrm{N}}(C, x)$ in (4.4), where $C \in \mathbb{R}^{1\times J}$. If Assumption A holds, then under $H_{0\mathrm{N}}(C, x)$ in (2.1), we have $t \xrightarrow{\mathrm{d}} \mathcal{N}(0, a)$ for some $a \in [0, 1]$. If Assumption B holds and $\pi \sim \mathrm{Unif}(\Pi_N)$, then $t_\pi | W \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1)$ a.s. In particular, the FRT with test statistic t_+ can asymptotically control type I error under $\tilde{H}_{0\mathrm{N}}(C, x)$ a.s.*

When $C \in \mathbb{R}^{m\times J}$ and $x \in \mathbb{R}^m$ for $m > 1$, we can interpret (4.4) as component-wise inequalities. Neither X^2 nor t_+ are acceptable when $m > 1$. An elementary workaround is to test each component using t_+ and apply a Bonferroni correction.

4.8 Cluster-Randomized Experiments

In many applied settings, the N units are partitioned into L clusters (e.g., classrooms in educational studies, villages in public health studies). All units belonging to a cluster must receive the same treatment. A cluster-randomized experiment assigns treatments to clusters, i.e. it is a CRE treating clusters as units. For $l = 1, \ldots, L$, let $\breve{W}_l \in \{1, \ldots, J\}$ represent the treatment that cluster l receives, and define the indicator $\breve{W}_l(j) = 1(\breve{W}_l = j)$. There are $L!/\prod_{j=1}^J L_j!$ possible realizations of $\{\breve{W}_1, \ldots, \breve{W}_L\}$. The mechanism of treatment assignment to clusters is identical to that to individuals in a CRE.

[66] stressed that we cannot implement the same analysis as if we had a CRE on the N units. For instance, $\hat{\bar{Y}}(j)$ is no longer an unbiased estimator for $\bar{Y}(j)$ if the cluster sizes

vary. Both [66] and [58] advised a CRE-like analysis. Let $X_i \in \{1, \ldots, L\}$ represent the cluster membership of unit i. Define cluster level aggregated potential outcomes $\{A_l(j) : l = 1, \ldots, L, j = 1, \ldots, J\}$, where $A_l(j) = \sum_{i=1}^{N} 1(X_i = l) Y_i(j)$. Define $A_l = (A_l(1), \ldots, A_l(J))^\top$, A_l^{obs}, $\bar{A} = (\bar{A}(1), \ldots, \bar{A}(J))^\top$, $\hat{\bar{A}} = (\hat{\bar{A}}(1), \ldots, \hat{\bar{A}}(J))^\top$ to align with our previous notation for a CRE. Aggregated potential outcomes resolve the problem of unbiased estimation of $\bar{Y}$: $\mathbb{E} L\hat{\bar{A}}/N = L\bar{A}/N = \bar{Y}$. Define $\hat{S}_A(j,j) = \sum_{l=1}^{L} \breve{W}_l(j)\{A_l^{\text{obs}} - \hat{\bar{A}}(j)\}^2/(L_j - 1)$ and $\hat{D}_A = L \cdot \text{diag}\{\hat{S}_A(1,1)/L_1, \ldots, \hat{S}_A(J,J)/L_J\}$. We revise the X^2 statistic as

$$X_A^2 = L(C\hat{\bar{A}} - Nx/L)^\top (C\hat{D}_A C^\top)^{-1} (C\hat{\bar{A}} - Nx/L).$$

Then Theorem 1 tells us that X_A^2 is proper for $H_{0\text{N}}(C, x)$ as $L \to \infty$ if Assumption B holds for the aggregated potential outcomes.

Chapter 5

Extensions

5.1 Stratified Randomized Experiments

We extend previous results to the stratified randomized experiment (SRE), also called the randomized block design. The overall setup from the CRE still applies, but now with each unit we also observe an associated covariate $X_i \in \{1, \ldots, H\}$. Thus, our data are $\{Y_i^{\text{obs}}, X_i, W_i : i = 1, \ldots, N\}$. The treatment does not affect this covariate. The W_i's remain the sole source of randomness. For $h = 1, \ldots, H$, the h-th stratum consists of all units i where $X_i = h$, with size $N_{[h]} = \sum_{i=1}^N 1(X_i = h)$ and proportion $\omega_{[h]} = N_{[h]}/N$. For $h = 1, \ldots, H$ and $j = 1, \ldots, J$, the experimenter predetermines the sample sizes $N_{[h]j} = \sum_{i=1}^N 1(X_i = h, W_i = j) \geq 2$. In a SRE, we assign treatments within each stratum just as we did in a CRE, and independently among different strata [46].

To define within-stratum means and covariances, we mirror previous notation. For $h = 1, \ldots, H$, the mean vector is $\bar{Y}_{[h]} \in \mathbb{R}^J$, which has j-th entry $\bar{Y}_{[h]}(j) = \sum_{i=1}^N 1(X_i = h)Y_i(j)/N_{[h]}$. The covariance $S_{[h]}$ has (j,k)-th entry $S_{[h]}(j,k) = \sum_{i=1}^N 1(X_i = h)\{Y_i(j) - \bar{Y}_{[h]}(j)\}\{Y_i(k) - \bar{Y}_{[h]}(k)\}/(N_{[h]} - 1)$. We stipulate the following regularity condition, which is in short for Assumption B to be true within all strata.

Assumption C. For $h = 1, \ldots, H$, (1) $\lim_{N\to\infty} N_{[h]}/N = \omega_{[h]} \geq 0$ and $\lim_{N\to\infty} N_{[h]j}/N_{[h]} = p_{[h]j} > 0$; (2) the sequences $(\bar{Y}_{[h]})$ and $(S_{[h]})$ converge to $\bar{Y}_{[h]\infty}$ and $S_{[h]\infty}$; (3) the matrix $S_{[h]\infty}$ has strictly positive main diagonal entries; (4) there exists an $L < \infty$ such that $\sum_{i=1}^N 1(X_i = h)\{Y_i(j) - \bar{Y}_{[h]}(j)\}^4/N_{[h]} \leq L$ for all N and $j = 1, \ldots, J$.

We do not distinguish between Assumptions A and B in the SRE for convenience. With a tiny abuse of notation, $\omega_{[h]}$ stands for both $N_{[h]}/N$ and its limit. The sample mean vector is $\hat{\bar{Y}}_{[h]} \in \mathbb{R}^J$, which has j-th entry $\hat{\bar{Y}}_{[h]}(j) = \sum_{i=1}^N 1(X_i = h, W_i = j)Y_i^{\text{obs}}/N_{[h]j}$. The sample variance is $\hat{S}_{[h]}(j,j) = \sum_{i=1}^N 1(X_i = h, W_i = j)\{Y_i^{\text{obs}} - \hat{\bar{Y}}_{[h]}(j)\}^2/(N_{[h]j} - 1)$. Under Assumption C, we have from Proposition 3 that, inside stratum h, the standardized stratum-wise sample mean $N_{[h]}^{1/2}(\hat{\bar{Y}}_{[h]} - \bar{Y}_{[h]})$ is asymptotically Normal with mean 0 and a covariance we denote $V_{[h]}$. A conservative estimator for $V_{[h]}$ is

$$\hat{D}_{[h]} = N_{[h]} \cdot \text{diag}\{\hat{S}_{[h]}(1,1)/N_{[h]1}, \ldots, \hat{S}_{[h]}(J,J)/N_{[h]J}\}.$$

An unbiased estimator for $\bar{Y}$ is $\breve{\bar{Y}} = \sum_{h=1}^{H} \omega_{[h]} \hat{\bar{Y}}_{[h]}$. Owing to the independence of treatment assignment across different strata, $N^{1/2}(\breve{\bar{Y}} - \bar{Y})$ is asymptotically Normal with mean 0 and covariance $\sum_{h=1}^{H} \omega_{[h]} V_{[h]}$. A conservative variance estimator is $\breve{D} = \sum_{h=1}^{H} \omega_{[h]} \hat{D}_{[h]}$.

We are now positioned to make an adjustment to X^2 that is proper when used with the FRT in a SRE:

$$\begin{aligned} X^2 &= N(C\breve{\bar{Y}} - x)^\top (C\breve{D}C^\top)^{-1}(C\breve{\bar{Y}} - x) \\ &= N \left(C \sum_{h=1}^{H} \omega_{[h]} \hat{\bar{Y}}_{[h]} - x \right)^\top \left(\sum_{h=1}^{H} \omega_{[h]} C \hat{D}_{[h]} C^\top \right)^{-1} \left(C \sum_{h=1}^{H} \omega_{[h]} \hat{\bar{Y}}_{[h]} - x \right) \end{aligned} \tag{5.1}$$

The special case $h = 1$ agrees with (3.1), so the same notation X^2 for this statistic is logical. Besides the form of the test statistic, the FRT entails two more modifications in the case of an SRE. First, we impute the potential outcomes stratum by stratum under the sharp null hypothesis

$$H_{0\mathrm{F}}(C, x_{[1]}, \ldots, x_{[H]}, \tilde{C}, \tilde{x}_{[1]}, \ldots, \tilde{x}_{[H]}) : \begin{pmatrix} C \\ \tilde{C} \end{pmatrix} Y_i^* = \begin{pmatrix} x_{[h]} \\ \tilde{x}_{[h]} \end{pmatrix}, \text{ whenever } X_i = h.$$

Since we still aim to test (2.1), the above null hypothesis must satisfy $\sum_{h=1}^{H} \omega_{[h]} x_{[h]} = x$. If $x = 0_m$, it is natural to choose $x_{[h]} = x$ and $\tilde{x}_{[h]} = 0_{J-m-1}$ for each h. Under the above sharp null hypothesis, we can impute all potential outcomes: for units with $X_i = h$,

$$Y_i^* = \begin{pmatrix} Y_i^*(1) \\ \vdots \\ Y_i^*(J) \end{pmatrix} = z_{[h]} + (Y_i^{\mathrm{obs}} - z_{[h],W_i}) 1_J, \text{ where } z_{[h]} = \begin{pmatrix} z_{[h],1} \\ \vdots \\ z_{[h],J} \end{pmatrix} = \begin{pmatrix} C \\ \tilde{C} \\ 1_J^\top \end{pmatrix}^{-1} \begin{pmatrix} x_{[h]} \\ \tilde{x}_{[h]} \\ 0 \end{pmatrix},$$

or, equivalently, $Y_i^*(j) = Y_i^{\mathrm{obs}} + z_{[h],j} - z_{[h],W_i}$. Second, we ought to permute the treatment indicators within strata, independently across strata. Let $\Pi_{N,\mathrm{bl}} \subseteq \Pi_N$ be all such permutations from a SRE. The p-value is $\left(\prod_{h=1}^{H} N_{[h]}! \right)^{-1} \sum_{\pi \in \Pi_{N,\mathrm{bl}}} 1(X_\pi^2 \geq X^2)$.

Theorem 4. *In a SRE, suppose Assumption C holds. Under $H_{0\mathrm{N}}(C, x)$, $X^2 \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2$, where each $a_j \in [0, 1]$. If $\pi \sim \mathrm{Unif}(\Pi_{N,\mathrm{bl}})$, then $X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_m^2$ a.s. In particular, the FRT with test statistic X^2 can asymptotically control type I error because the condition of Proposition 4 holds.*

Even if the original experiment is a CRE, if a discrete covariate X is available, we can condition on the number of treated and control units within all strata. Then the treatment assignment is identical to a SRE. Therefore, in a CRE, we can still permute the treatment indicators within each stratum of X. This plan is billed as a conditional randomization test. [95] and [40] perceived that conditional randomization tests typically enhance the power as long as the covariates are predictive of the outcomes. [43] and [67] discussed post-stratification, the estimation counterpart to testing.

We have focused on the SRE with large strata, i.e., $N_{[h]} \to \infty$ for $h \in 1, \ldots, H$, and H is fixed. Our theory does not cover SREs with many small strata, i.e., the $N_{[h]}$'s are bounded but $H \to \infty$ [31]. Although we conjecture that similar results hold in such cases, we defer technical details to future research.

5.2 Multiple Outcomes and Multiple Testings

We can lengthen the reach of our framework to the case where all potential outcomes $Y_i(j) \in \mathbb{R}^d$ are vectors. Define $\bar{Y}(j)$ and $\hat{\bar{Y}}(j) \in \mathbb{R}^d$ as before. It is convenient to gather these into long vectors

$$\bar{Y} = \begin{pmatrix} \bar{Y}(1) \\ \vdots \\ \bar{Y}(J) \end{pmatrix} \in \mathbb{R}^{dJ}, \qquad \hat{\bar{Y}} = \begin{pmatrix} \hat{\bar{Y}}(1) \\ \vdots \\ \hat{\bar{Y}}(J) \end{pmatrix} \in \mathbb{R}^{dJ}.$$

The covariances $S(j,k) = \sum_{i=1}^N \{Y_i(j) - \bar{Y}(j)\}\{Y_i(k) - \bar{Y}(k)\}^\top/(N-1)$ and $\hat{S}(j,j) = \sum_{i=1}^N W_i(j)\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}^\top/(N_j - 1)$ are now matrices, for $j,k = 1, \ldots, J$. The overall covariance matrix $S \in \mathbb{R}^{dJ \times dJ}$ has (j,k)-th block $S(j,k)$. Assume $S(j,j)$ and $\hat{S}(j,j)$ are both positive definite for all realizations of W.

Let $Y_i(j)_1, \ldots, Y_i(j)_d$ be the components of the potential outcomes $Y_i(j)$ for all i and j. We are interested in testing the weak null hypothesis

$$H_{0\text{N}}(C_1, \ldots, C_d, x_1, \ldots, x_d) : C_1 \begin{pmatrix} \bar{Y}(1)_1 \\ \vdots \\ \bar{Y}(J)_1 \end{pmatrix} = x_1, \ldots, C_d \begin{pmatrix} \bar{Y}(1)_d \\ \vdots \\ \bar{Y}(J)_d \end{pmatrix} = x_d, \tag{5.2}$$

where $C_1, \ldots, C_d$ are contrast matrices that have J columns and possibly varying row counts. We can condense notation with the Kronecker product: define

$$C = \begin{pmatrix} C_1 \otimes e_1^\top \\ \vdots \\ C_d \otimes e_d^\top \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix},$$

where $\{e_1, \ldots, e_d\}$ are the standard basis vectors of $\mathbb{R}^d$. We can write the above null hypothesis in the more compact form $H_{0\text{N}}(C, x) : C\bar{Y} = x$. It looks exactly like (2.1), but C cannot be an arbitrary contrast matrix.

Example 4. We lay out some possible contrast matrices when $J = 3$ and $d = 2$. The hypothesis $H_0 : \bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3)$ has the contrast matrix

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} C_1 \otimes e_1^\top \\ C_1 \otimes e_2^\top \end{pmatrix}, \text{ where } C_1 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}$$

Here, we test the same hypothesis entry by entry, and an equivalent contrast matrix is $C_1 \otimes I_2$. We can also test different hypotheses entry by entry, for instance $H_0 : \bar{Y}(1)_1 = \bar{Y}(2)_1$, $\bar{Y}(2)_2 = \bar{Y}(3)_2$. This hypothesis has the contrast matrix

$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} C_1 \otimes e_1^\top \\ C_2 \otimes e_2^\top \end{pmatrix}, \text{ where } C_1 = (1, -1, 0) \text{ and } C_2 = (0, 1, -1). \quad \square$$

The potential outcomes framework cannot withstand comparison of different entries under different treatments, for instance $H_0 : \bar{Y}(1)_1 = \bar{Y}(2)_2$. Null hypotheses like these do not have a clear causal interpretation here. Under iid sampling, [34] allow for a general contrast matrix C, and even for the length of $Y_i(j)$ to depend on treatment j. We constrain the contrast matrices C that we accept, as we have just detailed.

Under i.i.d. sampling and vector potential outcomes, [20] address the two-sample problem with permutation tests. [92], [54] and [35] test general linear hypotheses with bootstrap methods. We will use the FRT for (5.2). It is not a sharp null hypothesis, so we concoct one:

$$H_{0\mathrm{F}} : \begin{pmatrix} C_1 \\ \tilde{C}_1 \end{pmatrix} \begin{pmatrix} Y_i(1)_1 \\ \vdots \\ Y_i(J)_1 \end{pmatrix} = \begin{pmatrix} x_1 \\ \tilde{x}_1 \end{pmatrix}, \ldots, \begin{pmatrix} C_d \\ \tilde{C}_d \end{pmatrix} \begin{pmatrix} Y_i(1)_d \\ \vdots \\ Y_i(J)_d \end{pmatrix} = \begin{pmatrix} x_d \\ \tilde{x}_d \end{pmatrix}, \text{ for } i = 1, \ldots, N,$$

where the matrices $(C_1^\top, \tilde{C}_1^\top, 1_J)$ through $(C_d^\top, \tilde{C}_d^\top, 1_J)$ are invertible. We construct the $\tilde{C}$'s and $\tilde{x}$'s for each component of the outcome in the same way as the scalar case. In the hypothesis $H_{0\mathrm{F}}$, our notation does not reflect its dependence on the C's, $\tilde{C}$'s, x's and $\tilde{x}$'s. We impute potential outcomes as if $H_{0\mathrm{F}}$ were the reality. For the first component:

$$\begin{pmatrix} Y_i^*(1)_1 \\ \vdots \\ Y_i^*(J)_1 \end{pmatrix} = z_1 + (Y_{i,1}^{\mathrm{obs}} - z_{1W_i}) 1_J, \text{ where } z_1 = \begin{pmatrix} z_{11} \\ \vdots \\ z_{1J} \end{pmatrix} = \begin{pmatrix} C_1 \\ \tilde{C}_1 \\ 1_J^\top \end{pmatrix}^{-1} \begin{pmatrix} x_1 \\ \tilde{x}_1 \\ 0 \end{pmatrix} \tag{5.3}$$

and similarly for the second through the d-th entries, replacing all subscripts 1 by $2, \ldots, d$.

For vector potential outcomes, we tweak X^2 in (3.1):

$$X^2 = N(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1} (C\hat{\bar{Y}} - x),$$

where the block diagonal matrix $\hat{D} = N \cdot \mathrm{diag}\{\hat{S}(1,1)/N_1, \ldots, \hat{S}(J,J)/N_J\}$ is an asymptotically conservative estimator of $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$. This is in sync with (2.4). The FRT with X^2 can control the asymptotic type I error under (5.2). We first give the asymptotic requirements and then adapt Theorem 1 to the vector case. Let $|\cdot|$ be the Euclidean norm, which reduces to the usual absolute value for scalars.

Assumption D. The sequence (N_j/N) converges to $p_j \in (0,1)$ for all $j = 1, \ldots, J$. The sequences $(\bar{Y}_N)$ and (S_N) converge to $\bar{Y}_\infty$ and S_∞, where $|\bar{Y}_\infty| < \infty$, S_∞ is positive semi-definite, and $S_\infty(j,j)$ is positive definite for all $j = 1, \ldots, J$. Further,

$$\lim_{N \to \infty} \max_{j=1,\ldots,J} \max_{i=1,\ldots,N} |Y_i(j) - \bar{Y}(j)|^2 / N = 0.$$

Assumption E. Same as Assumption D with the last equation replaced by: there exists an $L < \infty$ such that $\sum_{i=1}^{N} |Y_i(j) - \bar{Y}(j)|^4/N \leq L$ for all $j = 1, \ldots, J$ and $N \geq (d+1)J$.

Proposition 7. *Assumption E implies Assumption D.*

Theorem 5. *If Assumption D holds, then under* $H_{0\mathrm{N}}(C, x)$, $X^2 \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2$, *where each* $a_j \in [0, 1]$. *If Assumption E holds and* $\pi \sim \mathrm{Unif}(\Pi_N)$, *then* $X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_m^2$ *a.s. In particular, the FRT with test statistic* X^2 *can asymptotically control type I error a.s.*

Theorem 5 puts in place a foundation for a single FRT for multiple outcomes. Following [20, Section 4], we can join Theorem 5 with the closure procedure for multiple testings. We omit the details.

To conduct the FRT with X^2 at all, we require all realizations of $\hat{S}(j, j)$ to be invertible, for which it is necessary that $N_j \geq d + 1$. [35] instead tried $\tilde{X}^2 = N(C\hat{\tilde{Y}} - x)^\top (C\tilde{D}C^\top)^{-1}(C\hat{\tilde{Y}} - x)$ with a bootstrap, where $\tilde{D}$ is a diagonal matrix with the same main diagonal as $\hat{D}$. However, $\tilde{X}^2$ is not proper for the FRT because the asymptotic distribution of $\tilde{X}_\pi^2 | W$ is not pivotal. So it is flawed for the same reason the Box type statistic B in (3.2) is. We reserve FRTs with $d \to \infty$ for future research.

Chapter 6

Simulations

6.1 Breaking the Box-Type Statistic

Previous sections show that X^2 is proper, but B and F are not. As a complement to this asymptotic fact, simulations reveal their finite sample behavior. To drive this point, we repeat the simulations with varying sample sizes. All the test statistics we brought up had other specific purposes in the literature. Thus, the simulations also serve to compare their efficacy with the FRT for testing weak null hypotheses.

We decided on the ANOVA with $J = 3$ treatment arms and a 2^2 Factorial ($J = 4$) setup, both with a balanced design $N_j = N/J$ for all j. We then gain from Proposition 5 that $B = F$. Thus, a comparison of X^2 and B suffices. In all cases, we force $\bar{Y}(1) = \cdots = \bar{Y}(J) = 0$, so the weak null hypothesis of no treatment effects on average holds. We also compel the potential outcomes to have covariance structure $S = uu^\top$. For the ANOVA case, $u^\top = (u_1, u_2, u_3) = (1, 2, 3)$, and for the Factorial case, $u^\top = (u_1, u_2, u_3, u_4) = (3, 1, 1, 3)$.

Explicitly, we first generate $Y_i(1) \overset{iid}{\sim} \mathcal{N}(0, 1)$ for $i = 1, \ldots, N$, center them, and scale them according to $Y_i(j) = u_j Y_i(1)$. For the hypothesis test itself, we simulate 10000 different realizations of the observed outcomes. For each set of $(W_i, Y_i^{\text{obs}})_{i=1}^N$, we run the FRT with both X^2 and B, calculating p-values from 2500 permutations.

For these potential outcomes, we perceive from Theorems 1 and 2 that the permutation distributions of X^2 and $2B$ are asymptotically χ^2_2 in both the ANOVA and factorial designs, but their asymptotic randomization distributions under $H_{0\text{N}}$ are

$$\begin{aligned} X^2 &\xrightarrow{\text{d}} \xi_1^2 + 0.758\xi_2^2, \qquad 2B \xrightarrow{\text{d}} 1.423\xi_1^2 + 0.434\xi_2^2, \qquad \text{(ANOVA)}, \\ X^2 &\xrightarrow{\text{d}} \xi_1^2 + \xi_2^2 \overset{\text{d}}{=} \chi^2_2, \qquad 2B \xrightarrow{\text{d}} 1.8\xi_1^2 + 0.2\xi_2^2, \qquad \text{(Factorial)}, \end{aligned} \tag{6.1}$$

providing an illustrative and simple numerical example of our main results. Each weight for X^2 is at most 1, while the weights for $2B$ are only at most 1 on average. In the Factorial case, the FRT with X^2 is actually asymptotically exact because both the randomization and permutation distributions of X^2 approach χ^2_2.

We can naturally broaden the ANOVA simulation just performed to SREs. We keep the ANOVA setup, but now incorporate a SRE with $H = 2$ strata. Remember that this means

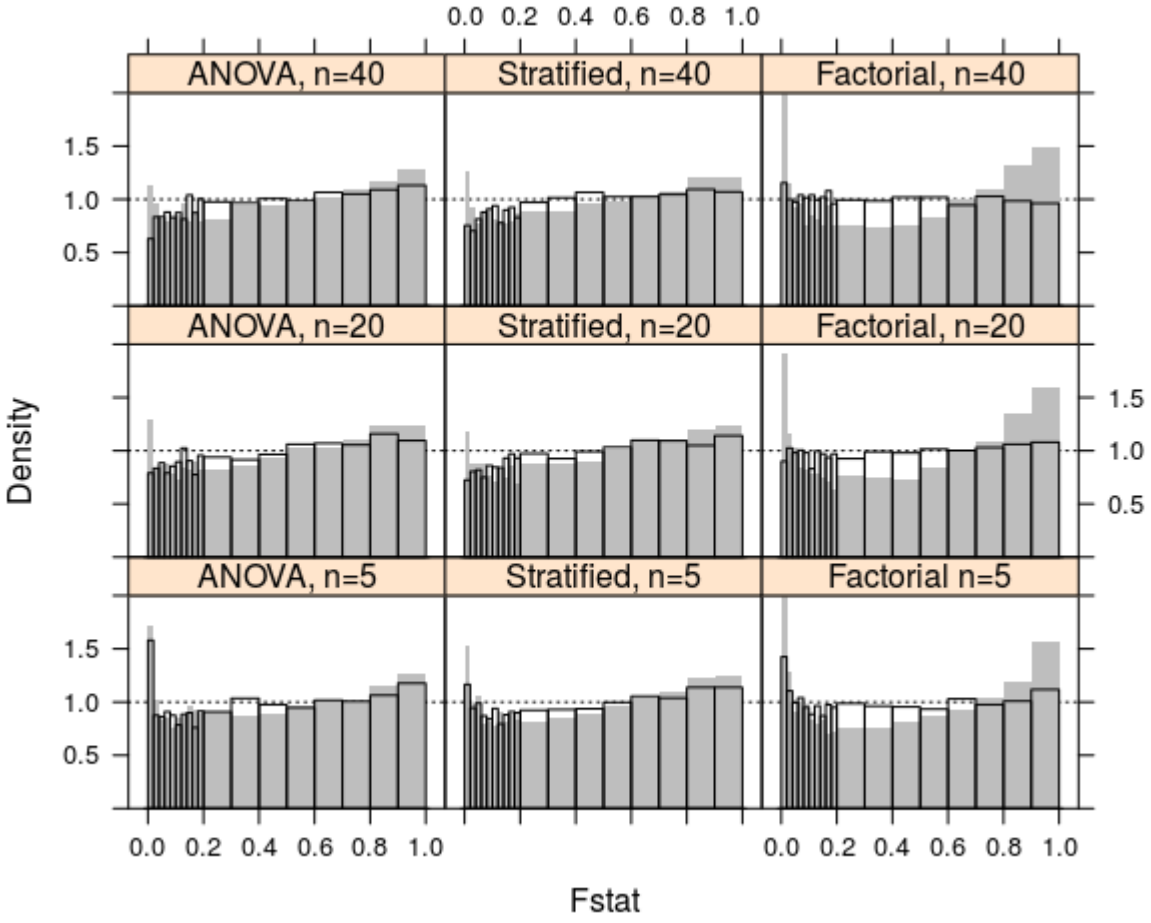

Figure 6.1: Histograms of FRT p-values under various settings and sample sizes. Gray bars indicate p-values from a F statistic, while transparent bars indicate p-values from the X^2 statistic. We display smaller p-values with a finer resolution because most hypothesis tests are conducted at levels close to 0. A dashed line indicating the Unif$(0, 1)$ density is added for reference purposes.

the observed data come from running a CRE within each stratum separately. The first stratum of potential outcomes will be identical to those of the ANOVA simulation above. The second stratum will be identical to the first, except with a unit constant added to all potential outcomes. This between stratum effect merits a SRE analysis. We proceed with the X^2 statistic in (5.1), and only permute data within each stratum when obtaining p-values.

The textbook suggestion [68] for testing the one-way ANOVA hypothesis in the SRE case involves the F statistic from a linear regression of the observed response on stratum and treatment indicators, ie $J + H$ predictors. Although [68] has reiterated the usual OLS assumptions that justify the F test, practitioners do not always check them. We therefore

would like to compare X^2 and F in this SRE setting. From Theorem 4, we know X^2 in (5.1) has the same asymptotic behavior as listed in (6.1). By intuition from [59], we anticipate that $2F$ also has the same asymptotic behavior as before.

In all three settings we have put forth, we also fix three different sample size settings to pinpoint the rate that asymptotics take effect. These are N_1 =5, 20, and 40 for ANOVA (without stratification) and Factorial, and the same counts for $N_{[1]1}$ for ANOVA with stratification.

Figure 6.1 contains the simulation results. For each setting and sample size, we plot histograms of p-values from the FRT with X^2 and B or F. In all histograms, the left-most bin of p-values ranging from 0 to 2% is most informative. For a successful control of type I error, the density of p-values here should not surpass 1 by much. From the bottom row of Figure 6.1, N_1 or $N_{[1]1} = 5$ (bottom row) is evidently far from the asymptotic regime. When N_1 or $N_{[1]1} = 20$ (middle row), it appears that we move much closer to the expected behavior dictated by asymptotics. This is because, when these counts are 40 (top row), the histograms do not change much.

It is also confirmed that the FRT with B or F fails to control type I error at small p-values for any sample size. We recollect from our theory that heteroscedasticity hampers its suitability. We have elected to balance the designs, so that it surfaces that, when $J > 2$, balanced designs do not guarantee the suitability of B or F as they do in treatment-control experiments (refer to Corollary 3). Of course, forgoing balanced designs can cause both B and F to fail more seriously. [24] compare X^2 and F in such cases through extensive simulation.

6.2 Confidence Regions

Our next simulation investigates constructing confidence regions alluded to by Corollary 4. At the same time, we seize the opportunity to compare the FRT and χ^2 approximations that are both asymptotically valid by Theorem 1. We decided on a balanced 2^2 factorial design ($K = 2$, $J = 2^2 = 4$) where $N_j = 10$ for $j = 1, \ldots, 4$. We seek to infer the main effects τ_1, τ_2, both individually and jointly. Take $Y_i(j) \overset{iid}{\sim} U^2 - 1/3$ where $U \sim \text{Unif}(0, 1)$, and center so that each $\bar{Y}(j) = 0$. This way, the weak null hypothesis of no treatment effects on average holds. Next, multiply each Y_i by the same matrix

$$\begin{pmatrix} 2 & 1 & 3/2 & 1 \\ 0 & \sqrt{5} & \sqrt{5}/2 & 2/\sqrt{5} \\ 0 & 0 & 3/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 0 & \sqrt{3.7} \end{pmatrix}$$

to inject correlation into the potential outcomes.

The set CR_α in (4.3) is a means to compute an asymptotic confidence region for τ_1, τ_2. Then we spread a grid of points centered at $\hat{\tau}_1$, $\hat{\tau}_2$ that comfortably encapsulates this asymptotic region. At each point (x_1, x_2) of this grid, we run the FRT with X^2 to test

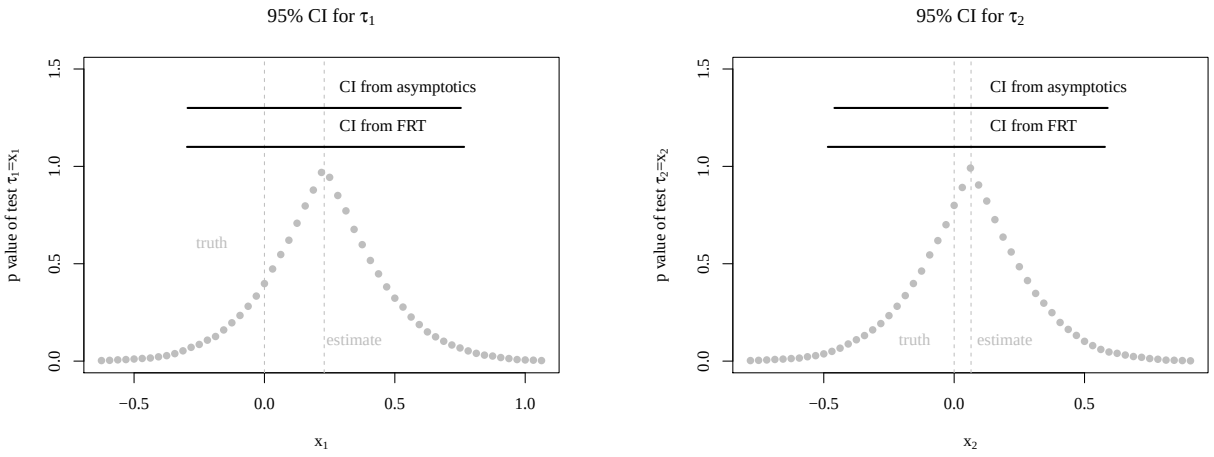

Figure 6.2: For τ_1 and τ_2 individually, the FRT and asymptotic approximation give nearly identical confidence intervals. For the second main effect, the FRT confidence interval is shifted due to the discrete resolution.

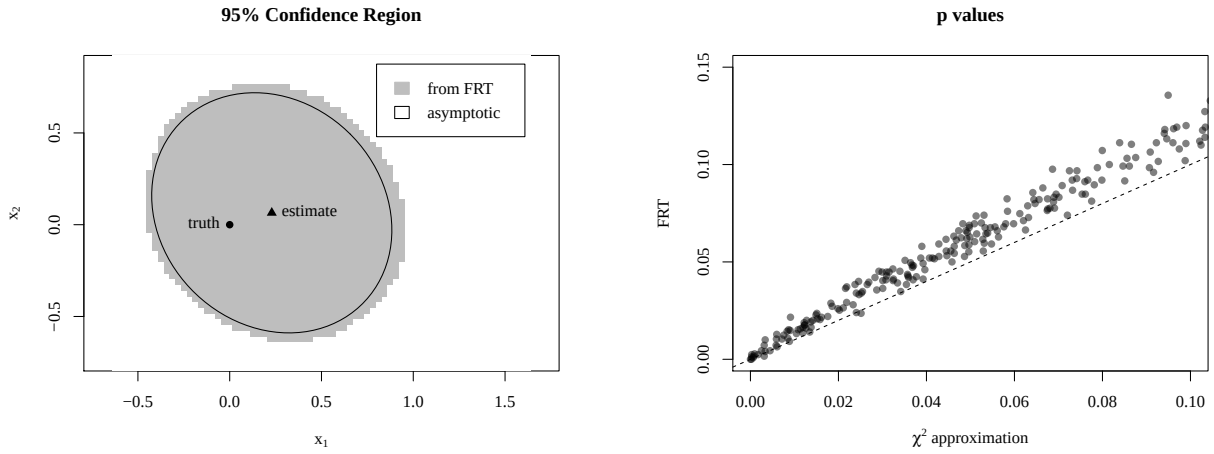

Figure 6.3: The left graph shows the FRT confidence region is again close to its asymptotic approximation, but the former is noticeably larger. The right graph is a scatter plot of p-values from testing $\tau_1 = \tau_2 = 0$ repeatedly from the original set of potential outcomes, zooming in on the region where they are less than 0.1.

$\tau_1 = x_1$, $\tau_2 = x_2$, both individually and jointly. We induct the point into our confidence region if and only if the p-value exceeds $\alpha = 0.05$.

Figure 6.2 shows the results for the marginal hypothesis tests. The behavior is very regular: the p-value crests near $\hat{\tau}_1$ or $\hat{\tau}_2$, and decays monotonically to the left and right. The FRT and χ^2 approximation confidence intervals are nearly indistinguishable.

Figure 6.3 shows the result for the joint test. The left graph shows the FRT confidence

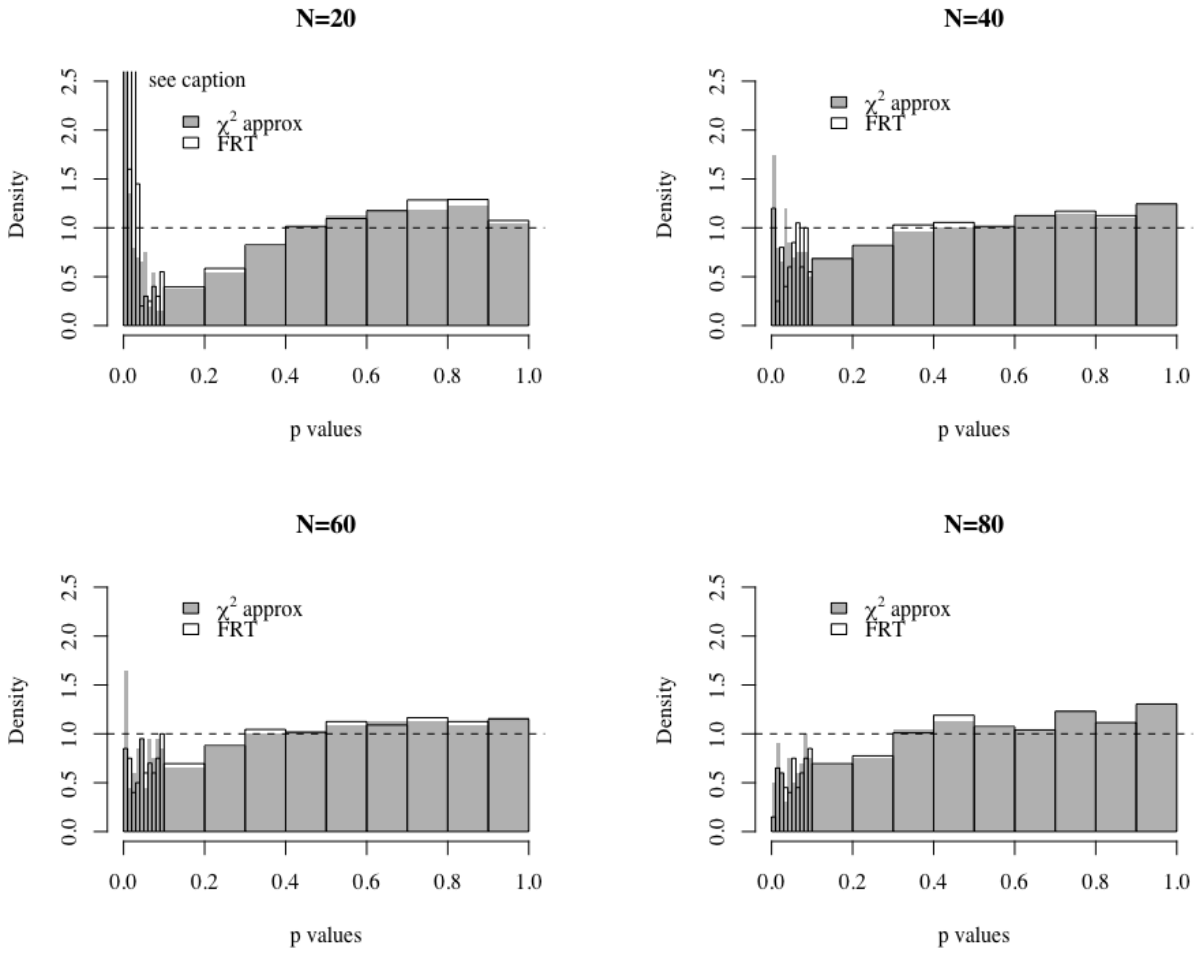

Figure 6.4: For N=20, the 0 to 0.01 bin density is 9.55 for the χ^2 approximation and 4.8 for the FRT. The 0.02 to 0.03 density is 2.75 for the FRT. For small samples, the tests do not perform as their asymptotics suggest.

region is again close to its asymptotic approximation, but not as close as in the 1D case. In particular, the former is noticeably larger. The right graph explains this by exposing that the p-values calculated from the FRT tend to be larger than those from the χ^2 approximation.

6.3 A treatment-control simulation

We can take advantage of the special setting of treatment-control to make the simulation run faster. The statistic X^2 in (3.1) reduces to

$$X^2 = \frac{\hat{\tau}^2}{\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2} \xrightarrow{\text{d}} a \cdot \chi^2_1, \text{ where } a = \frac{S(1,1)/p_1 + S(2,2)/p_2 - S^2_\tau}{S(1,1)/p_1 + S(2,2)/p_2}$$

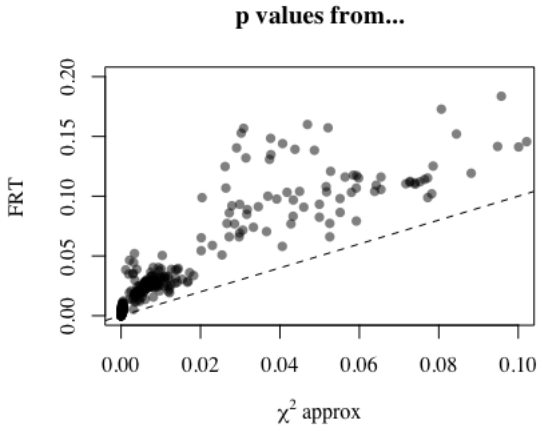

Figure 6.5: A scatter plot of p-values from the FRT versus those from the χ^2 approximation, when N=20. The dashed line on which the two p-values would be equal is included for reference.

Also recall $X^2_\pi | W \xrightarrow{\text{d}} \chi^2_1$ under $H_{0\text{N}}$. These claims are shown in the appendix. We agree that it is easier to calculate the above than (3.1). This leads to major savings in the time for the simulation because the FRT involves calculating X^2 repeatedly.

There are multiple goals for conducting this simulation. Recall that we have two methods for testing Neyman's null, both asymptotically conservative:

- (χ^2 approximation) Compute X^2 and compare it against the χ^2_1 distribution, to get a p-value. That is, the p-value is $\mathbb{P}(\chi^2_1 > X^2)$.
- Implement the FRT as described earlier.
- (Not possible in practice) Compare X^2 against the $a \cdot \chi^2_1$ distribution.

In [76], the permutation test (iid samples analog of FRT) is recommended. A major claim is that it empirically has better small sample properties than the χ^2 approximation. The goal is to investigate why this may be the case, as the theory is silent about finite sample performance.

The potential outcomes have to be generated somewhat pathologically in order to make the two tests perform noticeably differently. To create both $Y_i(1)$ and $Y_i(2)$, we generate 6 values iid from N(-2,1) and 14 values iid from N(2,1), and then the mean of the 20 values is subtracted off so that Neyman's Null holds exactly. The $Y_i(2)$ are further scaled by 2 so that the group variances are no longer approximately the same. Imbalanced designs are also needed to make the two tests perform differently. We chose $N_1 = 3N/4$. As in the previous section, permutation distributions were approximated by Monte Carlo using 2500 samples, and to simulate the behavior of X^2, 2000 realizations of treatment assignment were used.

Figure 6.4 shows the effects of increasing sample size in increments of 20, where the original 20 potential outcomes were repeated the necessary number of times. At N=20, both

tests fail to control type I error: the χ^2 approximation p-values are concentrated in the 0 to 0.01 bin, with a density of 9.55, and the FRT p-values are (slightly) more spread out between 0 to 0.04. However, this failure is mitigated with increasing sample size. Because of the failure to control type I error of the χ^2 approximation at all sample sizes except N=80, we note that knowledge of the constant a in (B.6) is often not helpful, since it would make the p-values even smaller.

Figure 6.5 is a scatter plot of p-values from the FRT versus those from the χ^2 approximation when N=20. Attention is restricted to p-values from the latter being < 0.1, as this is the most interesting region. Note that, most of the time, the p-value from the χ^2 approximation is smaller than that from the FRT, so the χ^2 approximation is overall more likely to result in a type I error, which illustrates the phenomenon alluded to in [76]. Figure 6.5 also shows that sometimes the two methods result in p-values that are quite far apart while other times the p-values are much closer together. Figure 6.6 investigates this in more detail. It displays the realization of the permutation distribution in each of those situations. In both cases of Figure 6.6, $X^2 \approx 9$. Yet, in the left graph, much mass is redistributed to make a fat right tail, increasing the FRT p-value, while in the right graph, the mass is redistributed to the region between 4 and 7. The right tail for the right graph is thus similar to that of the χ^2_1 distribution, and the permutation distribution does not help raise the p-value.

By looking at all the p-values, particularly those below 0.1, we build upon the simulations in [34], which focus on testing at significance level 0.05. Our simulations demonstrate that a test may be suitable for the level 0.05 (as the average of the densities in the bins from 0 to 0.05 is generally not large) but not at level 0.01, another common level. Of course there is nothing special about the p-values 0.05 and 0.01. A test may be fine for 0.01 but not, for instance, 0.001. The simulations seem to indicate that for very small observed p-values to be trustworthy, the sample size must be somewhat large. It is a quite regrettable reality that, albeit $X^2 \xrightarrow{\text{d}} a \cdot \chi^2_1$ under $H_{0\text{N}}$, the right tail of the finite sample distribution is often the most off from $a \cdot \chi^2_1$. This issue is compounded by the fact that experimenters tend to pick smaller levels to control type I error.

As a final concluding remark, treatment-control is the most elementary setting, but it still illuminates some unexpected results. Since more complicated settings like ANOVA or factorial designs build on treatment-control, it is reasonable to expect that all the findings from this simulation will also hold true in those settings.

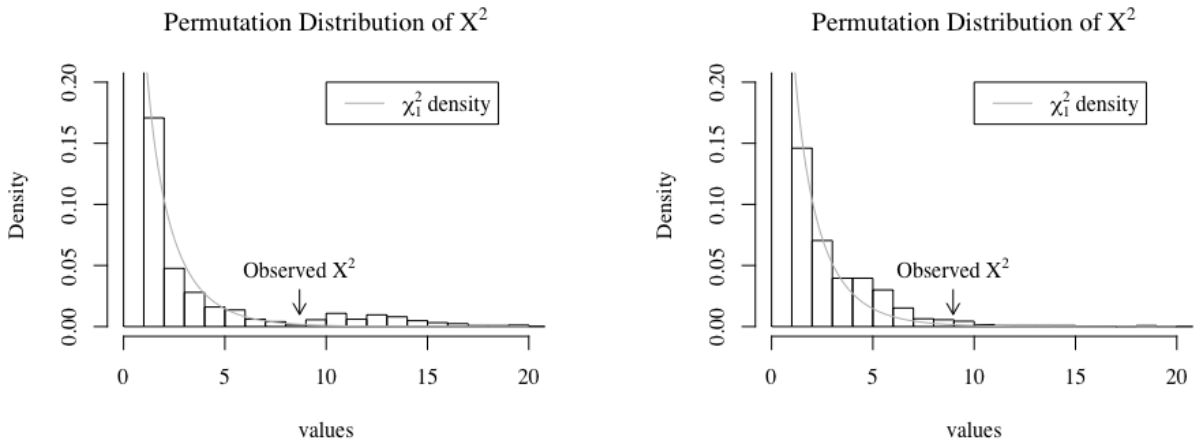

Figure 6.6: Two realizations of the permutation distribution, under different treatment assignments for potential outcomes B and N=20. In the first case, the p-value using the χ^2 approximation was < 0.01 while the p-value using the FRT was > 0.052. In the latter case, the p-values calculated from either method were both near 0.01.

Chapter 7

Applications and Discussion

We now try out our method on practical datasets, under a variety of possible weak null hypotheses. Our goal is not to do complete data analyses. We do not delve into issues of multiple comparisons. We pretend each null hypothesis is tested in isolation.

7.1 Financial Incentives for Exercise

[17] were interested in whether financial incentives caused college students to exercise more. They randomly assigned 40 students each to one of three possible treatments: no financial incentive (control), a small one, or a large one. We henceforth index these groups by $j = 1, 2, 3$, respectively. Then $N_1 = N_2 = N_3 = 40$. For each student, the response was the average number of weekly gym visits after the study minus that before the study. Let $Y_i(j)$ denote this quantity for the i-th student, if s/he received treatment j. Many students had $Y_i^{\text{obs}} = 0$. This would be problematic for the FRT with X^2 if, after a certain permutation, all permuted observations in a group were 0. To preclude this, we added a minuscule amount of random noise to all the Y_i^{obs}. For this dataset, the sample means are $-0.02937, 0.05414, 0.6398$, and the sample variances are $0.1523, 0.3859, 1.489$, for groups $j = 1, 2, 3$, respectively. Mere inspection of these numbers posits that a large financial incentive has a positive effect while a small one does not. It is also apparent that the data are heteroscedastic.

We test these four hypotheses at level 1% : whether the two magnitudes of financial incentives have any effect on average, whether financial incentives have any effect ignoring the distinction between large and small, whether financial incentives have any effect, and whether small financial incentives have any effect. In symbols, these are $2\bar{Y}(1) = \bar{Y}(2)+\bar{Y}(3)$, $\bar{Y}(1) = \bar{Y}(2, 3)$ (here we collapse treatment levels $j = 2, 3$ to one), $\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3)$, and $\bar{Y}(1) = \bar{Y}(2)$ (here we ignore the $j = 3$ observations), respectively.

We use the X^2 and F statistics, and get p-values both by the FRT and the χ^2 (or F) approximation. As we brought up earlier, p-values from FRTs are also finite-sample exact for testing Fisher's sharp null hypothesis. Consult Table 7.1 for the results. The class of hypothesis test (FRT and χ^2 (or F) approximation) holds little sway. It seems, for X^2, the FRT is slightly more conservative. For F, the FRT is slightly less conservative. Testing

Table 7.1: Analyzing [17]'s data. We report p-values as percents, and calculate the FRT p-values using 10^4 Monte Carlo simulations.

Hypothesis	$X^2 \overset{\text{d}}{\to} \chi^2_m$	FRT using X^2	$F \overset{\text{d}}{\to} F_{m,N-J}$	FRT using F
$2\bar{Y}(1) = \bar{Y}(2) + \bar{Y}(3)$	0.2522	0.27	1.970	1.59
$\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3)$	0.4189	0.49	0.06198	0.01
$\bar{Y}(1) = \bar{Y}(2,3)$	0.3353	0.49	2.454	2.34
$\bar{Y}(1) = \bar{Y}(2)$	47.15	47.93	47.37	47.93

the first two hypotheses, financial incentives have a statistically significant impact on gym attendance. Guided by Theorems 1 and 3, we should trust the p-values from X^2 more than those from F. The latter statistic seems to have overly conservative behavior for this dataset. Testing the third hypothesis suggests that the treated group ($j = 2$ or 3) has different behavior from the control in a statistically significant way.

With evidence that financial incentives might be helpful, we test the fourth hypothesis only comparing the control and small incentive groups, and get insignificant p-values. Note, in this case, $X^2 = F$ by Corollary 3, thanks to the balanced design. To wrap up, we concur with the findings of [17], that large financial incentives seem to induce people to visit the gym more often, but not small ones.

7.2 A 2^2 Factorial Experiment for Grades

We now undertake a similar analysis as in the previous section on another dataset. [1] wondered whether academic support services and/or financial incentives caused college students to improve their grades. Their data consisted of student grades for a certain semester on a 100 point scale. In that semester, students were either in a control group, offered a fellowship, offered services, or both. We thus have a 2^2 factorial experiment, and henceforth index these treatment groups by $j = 1, 2, 3, 4$, respectively. Unlike in the previous section, this experiment was imbalanced with $(N_1, N_2, N_3, N_4) = (854, 219, 212, 119)$. The sample means are $63.9, 65.8, 64.1, 66.1$, and the sample variances are $145, 124, 160, 114$, for groups $j = 1, 2, 3, 4$, respectively. By eye, there is less heteroscedasticity, and the sample means differ less markedly from the previous section.

We test the following five hypotheses at level 1%: financial services have no effect, services have no effect, neither has an effect, no interactions, and that all group means are the same. In symbols, these are $\bar{Y}(1) + \bar{Y}(2) = \bar{Y}(3) + \bar{Y}(4)$, $\bar{Y}(1) + \bar{Y}(3) = \bar{Y}(2) + \bar{Y}(4)$, both of the previous two, $\bar{Y}(1) + \bar{Y}(4) = \bar{Y}(2) + \bar{Y}(3)$, and $\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3) = \bar{Y}(4)$.

We again use the X^2 and F statistics, and get p-values both by the FRT and the χ^2 (or F) approximation. As we discussed earlier, p-values from FRTs are also exact for testing Fisher's sharp null hypothesis. Consult Table 7.2 for the results. The class of hypothesis test again holds little sway. The FRT seems overall slightly more conservative, but with some exceptions. We cannot reject any of these null hypotheses at level 1%. From the second and fourth hypotheses, the data do not seem to suggest services have any effect, or that there is

Table 7.2: Analyzing [1]'s data. We report p-values as percents, and calculate the FRT p-values using 10^4 Monte Carlo simulations.

Hypothesis	$X^2 \overset{\text{d}}{\to} \chi^2_m$	FRT using X^2	$F \overset{\text{d}}{\to} F_{m,N-J}$	FRT using F
No effect from services	72.84	72.34	73.92	73.58
No effect from incentives	1.192	1.43	1.602	1.80
No effects from either	3.652	3.99	5.262	5.28
No interaction	99.53	99.47	99.55	99.5
$\bar{Y}(1) = \bar{Y}(2) = \bar{Y}(3) = \bar{Y}(4)$	3.880	4.31	5.849	5.71

a non-additive effect from combining incentives and services. We do, however, almost reject the hypothesis of no effect from incentives alone, with p-values just over 1%.

Our finding that the effect of incentives is more significant than the effect of others conforms with the conclusions of [1]. They went on to conduct subgroup analysis, and discovered that the observed effects on grades come nearly exclusively from female students.

7.3 Discussion

We have proposed a strategy for using the FRT to test a weak null hypothesis. It imputes the missing potential outcomes under a compatible sharp null hypothesis, and then uses the studentized statistic in the FRT. It furthers the current literature in two directions. First, it complements the tests centered on asymptotic distributions. Our FRT is also finite-sample exact under the sharp null hypothesis. Second, it guides the choice of test statistic for the sharp null hypothesis. Although the finite-sample exactness property of the FRT holds for any test statistic, the p-values are sensitive to this choice. For example, all the p-values in Tables 7.1 and 7.2 are valid for Fisher's sharp null hypothesis. Unfortunately, these p-values range above and below the nominal significance level. This can be confusing in practice. Therefore, it is imperative to bring in weak null hypotheses and then studentized statistics. Our FRTs can control asymptotic type I error under weak null hypotheses and have power under corresponding alternative hypotheses.

Our theory ignores covariates. The analysis of covariance is a classical topic [30] and still attracts attention [59, 62, 32, 31, 65]. [10] and [56] widened it to the case where the number of covariates grows with the sample size. [93] and [80] discussed strategies for testing sharp null hypotheses. It is important to extend the theory to test weak null hypotheses with covariate adjustment, including the case with high dimensional covariates. We leave this to future work.

We have focused on completely randomized factorial experiments and extended the theory to stratified and clustered experiments. We conjecture that the strategy is also applicable for experiments with general treatment assignment mechanisms [69]. [33] also used the idea of studentization in sensitivity analysis of matched observational studies.

Bibliography

[1] J. Angrist, D. Lang, and P. Oropoulos. "Incentives and Services for College Achievement: Evidence from a Randomized Trial". In: *American Economic Journal: Applied Economics* 1 (2009), pp. 136–63.

[2] Joshua D Angrist and Jorn-Steffen Pischke. *Mostly Harmless Econometrics*. 1st ed. Princeton University Press, 2009.

[3] P. Armitage. "Tests for linear trends in proportions and frequencies". In: *Biometrics* 11.3 (1955), pp. 375–86.

[4] S. Athey, D. Eckles, and G. W. Imbens. "Exact P-values for Network Interference". In: *Journal of the American Statistical Association* 113 (2018), pp. 230–240.

[5] S. Athey and G. W. Imbens. "The Econometrics of Randomized Experiments". In: ed. by Esther Duflo Abhijit Banerjee. Vol. 1. Handbook of Economic Field Experiments. Elsevier B.V, 2017. Chap. 3, pp. 73–140.

[6] G. J. Babu and K. Singh. "Inference on means using the bootstrap". In: *The Annals of Statistics* 11 (1983), pp. 999–1003.

[7] G. Basse, A. Feller, and P. Toulis. "Exact tests for two-stage randomized designs in the presence of interference". In: *Biometrika* (2018), in press.

[8] D. Basu. "Randomization Analysis of Experimental Data: The Fisher Randomization Test". In: *Journal of the American Statistical Association* 75 (1980), pp. 575–582.

[9] R. L. Berger and D. D. Boos. "P values maximized over a confidence set for the nuisance parameter". In: *Journal of the American Statistical Association* 89 (1994), pp. 1012–6.

[10] A. Bloniarz et al. "Lasso adjustments of treatment effect estimates in randomized experiments". In: *Proceedings of the National Academy of Sciences of the United States of America* 113 (2016), pp. 7383–7390.

[11] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Non-asymptotic Theory of Independence*. Oxford University Press, 2016.

[12] G. Box. "Some theorems on quadratic forms applied in the study of analysis of variance problems". In: *Annals of Mathematical Statistics* 25 (1954), pp. 290–302.

[13] G. E. P. Box and S. L. Andersen. "Permutation theory in the derivation of robust criteria and the study of departures from assumption". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 17 (1955), pp. 1–34.

[14] J. V. Bradley. *Distribution-free statistical tests*. Upper Saddle River, NJ: Prentice Hall, 1968.

[15] E. Brunner, H. Dette, and A. Munk. "Box-Type Approximations in Nonparametric Factorial Designs". In: *Journal of the American Statistical Association* 92 (1997), pp. 1494–502.

[16] D. Caughey, A. Dafoe, and L. Miratrix. "Beyond the Sharp Null: Randomization Inference, Bounded Null Hypotheses, and Confidence Intervals for Maximum Effects". In: *arXiv preprint arXiv:1709.07339* (2017).

[17] G. Charness and U. Gneezy. "Incentives to Exercise". In: *Econometrica* 77 (2009), pp. 909–31.

[18] Ronald Christensen. *Plane Answers to Complex Questions*. 4th ed. Springer, 2011.

[19] E. Chung and J. Romano. "Exact and asymptotically robust permutation tests". In: *Annals of Statistics* 41 (2013), pp. 484–507.

[20] E. Chung and J. P. Romano. "Multivariate and multiple permutation tests". In: *Journal of Econometrics* 193.1 (2016), pp. 76–91.

[21] R. O. Collier and F. B. Baker. "Some Monte Carlo results on the power of the F-test under permutation in the simple randomized block design". In: *Biometrika* 53 (1966), pp. 199–203.

[22] T. Dasgupta, N. Pillai, and D. B. Rubin. "Causal inference from 2^K factorial designs by using potential outcomes". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 77 (2015), pp. 727–53.

[23] P. Ding. "A Paradox From Randomization-Based Causal Inference (with Discussion)". In: *Statistical Science* 32 (2017), pp. 331–45.

[24] P. Ding and T. Dasgupta. "A randomization-based perspective of analysis of variance: a test statistic robust to treatment effect heterogeneity". In: *Biometrika* 105 (2018), pp. 45–56.

[25] P. Ding, A. Feller, and L. Miratrix. "Randomization inference for treatment effect variation". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 78 (2016), pp. 655–671.

[26] E. Duflo, R. Glennerster, and M. Kremer. "Using randomization in development economics research: A toolkit". In: ed. by J. A. Strauss T. P. Schultz. Vol. 4. Elsevier, 2007. Chap. 61, pp. 3895–3962.

[27] T. Eden and F. Yates. "On the validity of Fisher's z test when applied to an actual example of non-normal data". In: *The Journal of Agricultural Science* 23 (1933), pp. 6–17.

[28] Julian J Faraway. *Linear models with R*. Chapman and Hall/CRC, 2016.

[29] R. A. Fisher. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

[30] R. A. Fisher. *The Design of Experiments*. 1st. Edinburgh, London: Oliver and Boyd, 1935.

[31] C. B. Fogarty. "On mitigating the analytical limitations of finely stratified experiments". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (2018), in press.

[32] C. B. Fogarty. "Regression assisted inference for the average treatment effect in paired experiments". In: *Biometrika* (2018), in press.

[33] C. B. Fogarty. "Studentized sensitivity analysis for the sample average treatment effect in paired observational studies". In: *arXiv preprint arXiv:1609.02112* (2016).

[34] S. Friedrich, E. Brunner, and M. Pauly. "Permuting longitudinal data in spite of the dependencies". In: *Journal of Multivariate Analysis* 153 (2017), pp. 255–265.

[35] S. Friedrich and M. Pauly. "MATS: Inference for potentially singular and heteroscedastic MANOVA". In: *Journal of Multivariate Analysis* 165 (2018), pp. 166–179.

[36] M. H. Gail et al. "On design considerations and randomization-based inference for community intervention trials". In: *Statistics in Medicine* 15 (1996), pp. 1069–1092.

[37] A. S. Gerber and D. P. Green. *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton and Company, 2012.

[38] P. Hall. "Theoretical comparison of bootstrap confidence intervals". In: *The Annals of Statistics* 16 (1988), pp. 927–953.

[39] David A. Harville. *Matrix Algebra From a Statistician's Perspective*. Springer, 1997.

[40] J. Hennessy et al. "A conditional randomization test to account for covariate imbalance in randomized experiments". In: *Journal of Causal Inference* 4 (2016), pp. 61–80.

[41] J. L. Hodges and E. L. Lehmann. "Estimates of Location Based on Rank Tests". In: *The Annals of Mathematical Statistics* 34 (1963), pp. 598–611.

[42] W. Hoeffding. "The large-sample power of tests based on permutations of observations". In: *The Annals of Mathematical Statistics* 23 (1952), pp. 169–192.

[43] D. Holt and T. M. F. Smith. "Post Stratification". In: *Journal of the Royal Statistical Society: Series A (General)* 142.1 (1979), pp. 33–46.

[44] P. J. Huber. "The behavior of maximum likelihood estimates under nonstandard conditions". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* 1 (1967), pp. 221–33.

[45] G. W. Imbens and K. Menzel. *A Causal Bootstrap*. Tech. rep. National Bureau of Economic Research, 2018.

[46] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press, 2015.

[47] Gareth James et al. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[48] A. Janssen. "Studentized permutation tests for non-iid hypotheses and the generalized Behrens-Fisher problem". In: *Statistics and Probability Letters* 36 (1997), pp. 9–21.

[49] A. Janssen. "Testing nonparametric statistical functionals with applications to rank tests". In: *Journal of Statistical Planning and Inference* 81 (1999), pp. 71–93.

[50] A. Janssen and T. Pauls. "How do bootstrap and permutation tests work?" In: *Annals of Statistics* 31 (2003), pp. 768–806.

[51] Robert W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer, 2010.

[52] O. Kempthorne. *The Design and Analysis of Experiments*. New York: John Wiley and Sons, 1952.

[53] O. Kempthorne and T. E. Doerfler. "The behaviour of some significance tests under experimental randomization". In: *Biometrika* 56 (1969), pp. 231–248.

[54] F. Konietschke et al. "Parametric and nonparametric bootstrap methods for general MANOVA". In: *Journal of Multivariate Analysis* 140 (2015), pp. 291–301.

[55] E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day, Inc., 1975.

[56] Lihua Lei and Peng Ding. "Regression adjustment in randomized experiments with a diverging number of covariates". In: *arXiv preprint arXiv:1806.07585* (2018).

[57] X. Li and P. Ding. "Exact confidence intervals for the average causal effect on a binary outcome". In: *Statistics in Medicine* 35 (2016), pp. 957–960.

[58] X. Li and P. Ding. "General forms of finite population central limit theorems with applications to causal inference". In: *Journal of the American Statistical Association* 112 (2017), pp. 1759–69.

[59] W. Lin. "Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique". In: *The Annals of Applied Statistics* 7 (2013), pp. 295–318.

[60] W. Lin et al. "A "placement of death" approach for studies of treatment effects on ICU length of stay". In: *Statistical Methods in Medical Research* 26 (2017), pp. 292–311.

[61] W. W. Loh, T. S. Richardson, and J. M. Robins. "An apparent paradox explained". In: *Statistical Science* 32 (2017), pp. 356–361.

[62] J. Lu. "Covariate adjustment in randomization-based causal inference for 2^K factorial designs". In: *Statistics and Probability Letters* 119 (2016), pp. 11–20.

[63] J. Lu. "On randomization-based and regression-based inferences for 2^K factorial designs". In: *Statistics and Probability Letters* 112 (2016), pp. 72–78.

[64] James G MacKinnon and Halbert White. "Some heteroskedasticity-Consistent Covariance Matrix estimators with improved finite sample properties". In: *Journal of Econometrics* 29.3 (1985), pp. 305–25.

[65] J. A. Middleton. "A Unified Theory of Regression Adjustment for Design-based Inference". In: *arXiv preprint arXiv:1803.06011* (2018).

[66] J. A. Middleton and P. M. Aronow. "Unbiased Estimation of the Average Treatment Effect in Cluster-Randomized Experiments". In: *Statistics, Politics and Policy* 6 (2015), pp. 39–75.

[67] L. Miratrix, J. Sekhon, and B. Yu. "Adjusting treatment effect estimates by post-stratification in Randomized Experiments". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 75 (2013), pp. 369–96.

[68] M. Morris. *Design of Experiments: An Introduction Based on Linear Models*. London: Chapman and Hall/CRC, 2010.

[69] R. Mukerjee, T. Dasgupta, and D. B. Rubin. "Using standard tools from finite population sampling to improve causal inference for complex experiments". In: *Journal of the American Statistical Association* (2018), in press.

[70] T. Mutze et al. "A studentized permutation test for three-arm trials in the gold standard design". In: *Statistics in Medicine* 36 (2017), pp. 883–98.

[71] G. Neuhaus. "Conditional rank tests for the two-sample problem under random censorship". In: *The Annals of Statistics* 21 (1993), pp. 1760–1779.

[72] J. Neyman. "On the Application of Probability Theory to Agricultural Experiments". In: *Statistical Science* 5 (1990), pp. 465–472.

[73] J. Neyman. "Statistical problems in agricultural experimentation (with discussion)". In: *Supplement to the Journal of the Royal Statistical Society* 2 (1935), pp. 107–180.

[74] T. L. Nolen and M. G. Hudgens. "Randomization-based inference within principal strata". In: *Journal of the American Statistical Association* 106 (2011), pp. 581–593.

[75] E. B. Page. "Ordered hypotheses for multiple treatments: a significance test for linear ranks". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 216–30.

[76] M. Pauly, E. Brunner, and F. Konietschke. "Asymptotic permutation tests in general factorial designs". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 77 (2015), pp. 461–73.

[77] E. J. G. Pitman. "Significance tests which may be applied to samples from any populations". In: *Supplement to the Journal of the Royal Statistical Society* 4 (1937), pp. 119–130.

[78] J. Rigdon and M. G. Hudgens. "Randomization inference for treatment effects on a binary outcome". In: *Statistics in Medicine* 34 (2015), pp. 924–935.

[79] J. P. Romano. "On the behavior of randomization tests without a group invariance assumption". In: *Journal of the American Statistical Association* 85 (1990), pp. 686–92.

[80] P. R. Rosenbaum. "Covariance adjustment in randomized experiments and observational studies". In: *Statistical Science* 17 (2002), pp. 286–327.

[81] P. R. Rosenbaum. "Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot". In: *Biometrika* 88 (2001), pp. 219–231.

[82] P. R. Rosenbaum. "Exact Confidence Intervals for Nonconstant Effects by Inverting the Signed Rank Test". In: *The American Statistician* 57 (May 2003), pp. 132–138.

[83] P. R. Rosenbaum. *Observational Studies*. 2nd ed. New York: Springer, 2002.

[84] P. R. Rosenbaum. "Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects". In: *Biometrics* 55 (1999), pp. 560–564.

[85] D. B. Rubin. "Causal Inference Using Potential Outcomes: Design, Modeling, Decisions". In: *Journal of the American Statistical Association* 100 (2005), pp. 322–31.

[86] D. B. Rubin. "Comment on D. Basu". In: *Journal of the American Statistical Association* 75 (1980), pp. 591–593.

[87] D. B. Rubin. "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66 (1974), pp. 688–701.

[88] A. Sabbaghi and D. B. Rubin. "Comments on the Neyman–Fisher controversy and its consequences". In: *Statistical Science* 29 (2014), pp. 267–284.

[89] C. Samii and P. M. Aronow. "On equivalencies between design-based and regression-based variance estimators for randomized experiments". In: *Statistics and Probability Letters* 80 (2012), pp. 365–70.

[90] J. Schmid. "A Remark on Characteristic Polynomials". In: *The American Mathematical Monthly* 77 (1970), pp. 998–9.

[91] P. Z. Schochet. "Multi-armed RCTs: A design-based framework". In: *Journal of Educational and Behavioral Statistics* (2018), in press.

[92] M. S. Srivastava and T. Kubokawa. "Tests for multivariate analysis of variance in high dimension under non-normality". In: *Journal of Multivariate Analysis* 115 (2013), pp. 204–216.

[93] J. W. Tukey. "Tightening the clinical trial". In: *Controlled Clinical Trials* 14 (1993), pp. 266–285.

[94] H. White. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". In: *Econometrica* 48 (1980), pp. 817–38.

[95] L. Zheng and M. Zelen. "Multi-center clinical trials: Randomization and ancillary statistics". In: *The Annals of Applied Statistics* 2 (2008), pp. 582–600.

Appendix A

Proofs and Lemmas for the Main Text

This first appendix is concerned with proving the results of the main text. In order to do so, we will also need some lemmas. Let us first review the existing notation and introduce some additional notation used in the appendix.

Let $X_N \xrightarrow{\text{d}} X$, $X_N \xrightarrow{\text{as}} X$ and $X_N \xrightarrow{P} X$ denote convergence in distribution, almost surely (often abbreviated "a.s."), and in probability, respectively. For random vectors or matrices, we use the same notation to denote such convergence, entry by entry. For convergence in probability, we may also write $\text{plim}_{N\to\infty} X_N = X$. Let $\xi_1, \xi_2, \ldots \overset{iid}{\sim} \mathcal{N}(0,1)$. For a diagonalizable matrix A, let $\lambda_j(A)$ be its j-th largest eigenvalue. Let $|\cdot|$ be the absolute value of a scalar or the Euclidean norm of a vector. Let $\|\cdot\|_F$ be the Frobenius norm of a matrix. For $A, B \in \mathbb{R}^{m\times n}$, let $A * B$ be the component-wise product of A and B: $(AB)_{ij} = A_{ij}B_{ij}$. Let $\max_i$, $\max_j$, and $\max_{i,j}$ denote the maximums over $\{i = 1, \ldots, n\}$, $\{j = 1, \ldots, J\}$, and both. Let $a \vee b = \max(a, b)$ be the maximum value of a and b. For any matrix $A \in \mathbb{R}^{m\times n}$, we let $A^- \in \mathbb{R}^{n\times m}$ denote any matrix satisfying $AA^-A = A$ and call A^- a *generalized inverse*. We let A^+ denote the (unique) particular generalized inverse that additionally satisfies $A^+AA^+ = A$ and AA^+, A^+A are symmetric, and call A^+ the *Moore-Penrose pseudoinverse*. "Simple random sample" will be abbreviated SRS.

A.1 Technical Lemmas

We require some linear algebra facts from [24].

Lemma 1. *(i) If $X \sim \mathcal{N}(0_J, A)$, then $X^\top BX \overset{\text{d}}{=} \sum_{j=1}^J \lambda_j(AB)\xi_j^2$. If A is a projection matrix, then each $\lambda_j(AB) \le \lambda_1(B)$.*

(ii) If $A, B \succeq 0$ and B is a correlation matrix, then $\lambda_1(A * B) \le \lambda_1(A)$.

(iii) If $X_n \xrightarrow{\text{d}} \mathcal{N}(0_m, A)$, and $B_n \xrightarrow{P} B \succ 0$, then $X_n^\top B_n^{-1} X_n \xrightarrow{\text{d}} \sum_{j=1}^m \lambda_j(AB^{-1})\xi_j^2$. If $B \succeq A$, then each $\lambda_j(AB^{-1}) \in [0, 1]$.

Proof. (i) and (ii) come from [24]. We prove (iii). By two applications of Lemma 7, we have $B_n^{-1} \xrightarrow{P} B^{-1}$ and $X_n^\top B_n^{-1} X_n \xrightarrow{\mathrm{d}} X^\top B^{-1} X$. By (i), $X^\top B^{-1} X \overset{\mathrm{d}}{=} \sum_{j=1}^m \lambda_j(AB^{-1})\xi_j^2$. If $B \succeq A$, then each $\lambda_j(AB^{-1}) \in [0,1]$ by the last item of Lemma 8. □

Lemma 2 (Massart Concentration Inequality)**.** *A population $(Y_1, \ldots, Y_N)$ has mean $\bar{Y}$ and variance $S = \sum_{i=1}^N (Y_i - \bar{Y})^2/(N-1)$. Let $A \subseteq \{1, \ldots, N\}$ be a SRS of size N_1, $\hat{\bar{Y}} = N_1^{-1} \sum_{i \in A} Y_i$, and $p_1 = N_1/N$. Then for $t \geq 0$*

$$\mathbb{P}(\hat{\bar{Y}} - \bar{Y} \geq t) \vee \mathbb{P}(\hat{\bar{Y}} - \bar{Y} \leq -t) \leq \exp\big(\frac{-35N_1^2}{36NS} t^2\big) = \exp\big(\frac{-35Np_1^2}{36S} t^2\big).$$

Lemma 2 is crucial for our proof of almost sure convergence for sampling without replacement, as we are about to see.

Lemma 3. *Let $\big(\{Y_{N,i} : i = 1, \ldots, N\}$ be a sequence of populations with means $(\bar{Y}_N)$ and variances (S_N). Suppose we take a simple random sample from each population of size $N_1 \geq 2$ with sample mean $\hat{\bar{Y}}_N$ and variance $\hat{S}_N$. Assume $\lim_{N\to\infty} N_1/N = p_1 > 0$.*

(i) If the sequence (S_N) is bounded above by $S_{\max} < \infty$, then $|\hat{\bar{Y}}_N - \bar{Y}_N| \xrightarrow{\mathrm{as}} 0$. If we also have $\lim_{N\to\infty} \bar{Y}_N = \bar{Y}_\infty$, then $\hat{\bar{Y}}_N \xrightarrow{\mathrm{as}} \bar{Y}_\infty$. Assumption A implies these results.

(ii) If there is $L < \infty$ such that $\sum_{i=1}^N (Y_{N,i} - \bar{Y}_N)^4/N \leq L$ for all N, then $|\hat{S}_N - S_N| \xrightarrow{\mathrm{as}} 0$. If we also have $\lim_{N\to\infty} S_N = S_\infty$, then $\hat{S}_N \xrightarrow{\mathrm{as}} S_\infty$. Assumption B implies these results.

Proof. (i) Because $p_{N,1} = N_1/N \to p_1$, we can pick a positive integer N^* such that $N \geq N^*$ implies $p_{N,1} > p_1/2$. Then by Lemma 2, there is a universal constant $C \in (0, \infty)$, independent of N, such that, for $N \geq N^*$ and $t \geq 0$,

$$\mathbb{P}(|\hat{\bar{Y}}_N - \bar{Y}_N| \geq t) \leq 2\exp\left\{-\frac{Np_{N,1}^2}{CS_N} t^2\right\} \leq 2\exp\left\{-\frac{p_1^2}{4CS_{\max}} Nt^2\right\}$$
$$\implies \sum_{N \geq N^*} \mathbb{P}(|\hat{\bar{Y}}_N - \bar{Y}_N| \geq t) \leq 2 \sum_{N \geq N^*} \exp\left\{-\frac{p_1^2}{4CS_{\max}} Nt^2\right\} < \infty.$$

By the Borel–Cantelli Lemma, $|\hat{\bar{Y}}_N - \bar{Y}_N| \xrightarrow{\mathrm{as}} 0$.

(ii) First, by the Cauchy-Schwarz Inequality, we have that for all N

$$S_N = \frac{1}{N-1} \sum_{i=1}^N (Y_{N,i} - \bar{Y}_N)^2 \leq \frac{N^{1/2}}{N-1} \left\{\frac{1}{N} \sum_{i=1}^N (Y_{N,i} - \bar{Y}_N)^4\right\}^{1/2} \leq \frac{N}{N-1} L^{1/2},$$

which is bounded above as $N \to \infty$, so by (i), $|\hat{\bar{Y}}_N - \bar{Y}_N| \xrightarrow{\mathrm{as}} 0$.

Second, let $W_{N,i}$ be the indicator for Y_i being in the simple random sample. Define as an intermediate quantity $\tilde{S}_N = \sum_{i=1}^N W_{N,i}(Y_{N,i} - \bar{Y}_N)^2/(N_1 - 1)$, which differs from $\hat{S}_N$ by an

almost surely zero quantity as $N \to \infty$:

$$\begin{aligned}
\hat{S}_N - \tilde{S}_N =& \frac{1}{N_1 - 1} \sum_{i=1}^{N} W_{N,i} \big\{ (Y_{N,i} - \hat{\bar{Y}}_N)^2 - (Y_{N,i} - \bar{Y}_N)^2 \big\} \\
=& \frac{1}{N_1 - 1} \big\{ 2(\bar{Y}_N - \hat{\bar{Y}}_N) \sum_{i=1}^{N} W_{N,i} Y_{N,i} + N_1 ((\hat{\bar{Y}}_N)^2 - \bar{Y}_N^2) \big\} \\
=& \frac{N_1}{N_1 - 1} \big\{ 2(\bar{Y}_N - \hat{\bar{Y}}_N) \hat{\bar{Y}}_N + (\hat{\bar{Y}}_N)^2 - \bar{Y}_N^2 \big\} \\
=& \frac{-N_1}{N_1 - 1} (\hat{\bar{Y}}_N - \bar{Y}_N)^2 \overset{\text{as}}{\to} 0.
\end{aligned}$$

Third, we note that the variance is bounded above for all N:

$$\mathrm{Var}\{(Y_{N,i} - \bar{Y}_N)^2\}_{i=1}^{N} \leq \frac{1}{N-1} \sum_{i=1}^{N} (Y_{N,i} - \bar{Y}_N)^4 \leq \frac{N}{N-1} L.$$

So by (i), $\big| \sum_{i=1}^{N} W_{N,i} (Y_{N,i} - \bar{Y}_N)^2 / N_1 - \sum_{i=1}^{N} (Y_{N,i} - \bar{Y}_N)^2 / N \big| \overset{\text{as}}{\to} 0$, and therefore

$$\begin{aligned}
|\tilde{S}_N - S_N| =& \left| \frac{N_1}{N_1 - 1} \frac{1}{N_1} \sum_{i=1}^{N} W_{N,i} (Y_{N,i} - \bar{Y}_N)^2 - \frac{N_1}{N_1 - 1} \frac{1}{N} \sum_{i=1}^{N} (Y_{N,i} - \bar{Y}_N)^2 \right. \\
& \left. + \frac{N - N_1}{(N-1)(N_1 - 1)} \frac{N-1}{N} S_N \right| \\
\leq& \frac{N_1}{N_1 - 1} \left| \frac{1}{N_1} \sum_{i=1}^{N} W_{N,i} (Y_{N,i} - \bar{Y}_N)^2 - \frac{1}{N} \sum_{i=1}^{N} (Y_{N,i} - \bar{Y}_N)^2 \right| \\
& + \frac{N - N_1}{(N-1)(N_1 - 1)} \frac{N-1}{N} S_N \\
\leq& \frac{N_1}{N_1 - 1} \left| \frac{1}{N_1} \sum_{i=1}^{N} W_{N,i} (Y_{N,i} - \bar{Y}_N)^2 - \frac{1}{N} \sum_{i=1}^{N} (Y_{N,i} - \bar{Y}_N)^2 \right| + \frac{1}{N_1 - 1} L^{1/2} \to 0.
\end{aligned}$$

We now finally have $|\hat{S}_N - S_N| \leq |\hat{S}_N - \tilde{S}_N| + |\tilde{S}_N - S_N| \overset{\text{as}}{\to} 0$. □

Lemma 4. *Under Assumption A and for all sequences of W, the imputed potential outcomes in FRT-2 satisfy* $\lim_{N \to \infty} \max_{i,j} \{Y_i^*(j) - \bar{Y}^*(j)\}^2 / N = 0$.

Proof. For convenience, let $\max_i$, $\max_j$, and $\max_{i,j}$ denote the max over $i = 1, \ldots, N$, $j = 1, \ldots, J$, and both. Because $(\bar{Y}(j))$ converges for all $j = 1, \ldots, J$, and the z_j's do not depend on N, we may pick $Y_{\max} \in \mathbb{R}$ such that for all N,

$$\max_j |\bar{Y}(j)| \vee \max_j |z_j - \bar{z}| \leq Y_{\max}.$$

Put $L_N = \max_{i,j}\{Y_i(j) - \bar{Y}(j)\}^2$, which is $o(N)$ by Assumption A. Then

$$\max_{i,j} |Y_i(j) - \bar{Y}(j)| = \big[\max_{i,j}\{Y_i(j) - \bar{Y}(j)\}^2\big]^{1/2} \le L_N^{1/2}.$$

Next,

$$\max_i |Y_i^{\text{obs}}| \le \max_{i,j} |Y_i(j)| \le \max_{i,j} |Y_i(j) - \bar{Y}(j)| + \max_j |\bar{Y}(j)| \le L_N^{1/2} + Y_{\max}.$$

We bound the magnitude of $\bar{Y}_\bullet^{\text{obs}} = \sum_{i=1}^N Y_i^{\text{obs}}/N$:

$$|\bar{Y}_\bullet^{\text{obs}}| \le \max_i |Y_i^{\text{obs}}| \le L_N^{1/2} + Y_{\max}, \quad \max_i |Y_i^{\text{obs}} - \bar{Y}_\bullet^{\text{obs}}| \le \max_i |Y_i^{\text{obs}}| + |\bar{Y}_\bullet^{\text{obs}}| \le 2(L_N^{1/2} + Y_{\max}).$$

Using the above bounds and the additional bound $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\max_i (Y_i^{\text{obs}} - \bar{Y}_\bullet^{\text{obs}})^2 = \big(\max_i |Y_i^{\text{obs}} - \bar{Y}_\bullet^{\text{obs}}|\big)^2 \le 4(L_N^{1/2} + Y_{\max})^2 \le 8(L_N + Y_{\max}^2).$$

Incorporating the z's, we have

$$\begin{aligned}
\max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2 &= \max_i (Y_i^{\text{obs}} - z_{W_i} - \bar{Y}_\bullet^{\text{obs}} + \bar{z})^2 \\
&\le 2\big\{\max_i (Y_i^{\text{obs}} - \bar{Y}_\bullet^{\text{obs}})^2 + \max_i (z_{W_i} - \bar{z})^2\big\} \\
&\le 16(L_N + Y_{\max}^2) + 2Y_{\max}^2,
\end{aligned}$$

which is $o(N)$ as desired. □

Note it is in Lemma 4 where we need $(\bar{Y}_N)$ to be bounded above in norm. It seems that we need for $\tilde{x}$ to approximate $\tilde{C}\bar{Y}$ to $o(N)$, and a convenient way to achieve this is to assume bounded above.

Now we visit the vector versions of Lemmas 3 and 4.

Lemma 5. *Let $\big(\{Y_{N,i} : i = 1, \ldots, N\}$ be a sequence of populations with means $\bar{Y}_N) \in \mathbb{R}^d$ and covariances S_N. Suppose we take a simple random sample from each population of size $N_1 \ge d+1$ with sample mean $\hat{\bar{Y}}_N$ and covariance $\hat{S}_N$. Assume $\lim_{N\to\infty} N_1/N = p_1 > 0$.*

(i) If the sequence $(\|S_N\|_F)$ is bounded above by $S_{\max} < \infty$, then $|\hat{\bar{Y}}_N - \bar{Y}_N| \xrightarrow{\text{as}} 0$. If we also have $\lim_{N\to\infty} \bar{Y}_N = \bar{Y}_\infty$, then $\hat{\bar{Y}}_N \xrightarrow{\text{as}} \bar{Y}_\infty$. Assumption D implies these results.

(ii) If there is $L < \infty$ such that $\sum_{i=1}^N |Y_{N,i} - \bar{Y}_N|^4/N \le L$ for all N, then $\|\hat{S}_N - S_N\|_F \xrightarrow{\text{as}} 0$. If we also have $\lim_{N\to\infty} S_N = S_\infty$, then $\hat{S}_N \xrightarrow{\text{as}} S_\infty$. Assumption E implies these results.

Proof. (i) Note that each component of Y_{Ni} meets Lemma 3, so $|\hat{\bar{Y}}_N - \bar{Y}_N| \xrightarrow{\text{as}} 0$ holds component by component.

(ii) Because each component of Y_{Ni} meets Lemma 3, each entry on the main diagonal of $\hat{S}_N - S_N$ converges almost surely to 0. It is thus enough to show convergence of the $(1,2)$th

entry, for then identical logic will show convergence of an arbitrary off-diagonal entry. Let Y_{1Ni} and Y_{2Ni} be the first and second entries of Y_{Ni}.

We follow the steps of Lemma 3 closely. First, $\|S_N\|_F$ is bounded above:

$$\begin{aligned}\|S_N\|_F =& \frac{1}{N-1}\Big\|\sum_{i=1}^{N}(Y_{Ni}-\bar{Y}_N)(Y_{Ni}-\bar{Y}_N)^\top\Big\|_F \\ \leq& \frac{1}{N-1}\sum_{i=1}^{N}|Y_{Ni}-\bar{Y}_N|^2 \leq \frac{N^{1/2}}{N-1}\left(\sum_{i=1}^{N}|Y_{Ni}-\bar{Y}_N|^4\right)^{1/2} \leq \frac{NL^{1/2}}{N-1},\end{aligned}$$

where the first inequality follows from the Triangle Inequality and $\|ab^\top\|_F = |a|\cdot|b|$ for two vectors a and b, and the second inequality by the Cauchy–Schwarz Inequality. By (i), $|\hat{\bar{Y}}_N - \bar{Y}_N| \xrightarrow{\text{as}} 0$.

Second, let $W_{N,i}$ be the indicator for Y_i being in the simple random sample. Define as an intermediate quantity $\tilde{S}_{12N} = \sum_{i=1}^{N} W_{Ni}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})/(N_1-1)$, which differs from $\hat{S}_{12N}$ by an almost surely zero quantity as $N\to\infty$:

$$\begin{aligned}\hat{S}_{12N}-\tilde{S}_{12N} =& \frac{1}{N_1-1}\sum_{i=1}^{N} W_{Ni}\{(Y_{1Ni}-\hat{\bar{Y}}_{1N})(Y_{2Ni}-\hat{\bar{Y}}_{2N}) - (Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\} \\ =& \frac{1}{N_1-1}\sum_{i=1}^{N} W_{Ni}\{(\bar{Y}_{1N}-\hat{\bar{Y}}_{1N})Y_{2Ni} + (\bar{Y}_{2N}-\hat{\bar{Y}}_{2N})Y_{1Ni} + \hat{\bar{Y}}_{1N}\hat{\bar{Y}}_{2N} - \bar{Y}_{1N}\bar{Y}_{2N}\} \\ =& \frac{N_1}{N_1-1}\{(\bar{Y}_{1N}-\hat{\bar{Y}}_{1N})\hat{\bar{Y}}_{2N} + (\bar{Y}_{2N}-\hat{\bar{Y}}_{2N})\hat{\bar{Y}}_{1N} + \hat{\bar{Y}}_{1N}\hat{\bar{Y}}_{2N} - \bar{Y}_{1N}\bar{Y}_{2N}\} \\ =& \frac{-N_1}{N_1-1}(\bar{Y}_{1N}-\hat{\bar{Y}}_{1N})(\bar{Y}_{2N}-\hat{\bar{Y}}_{2N}) \xrightarrow{\text{as}} 0.\end{aligned}$$

Third, we note that the variance is bounded above for all N:

$$\begin{aligned}\text{Var}\{(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})\}_{i=1}^{N} \leq& \frac{1}{N-1}\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})^2(Y_{2Ni}-\bar{Y}_{2N})^2 \\ \leq& \frac{1}{N-1}\left\{\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})^4\sum_{i=1}^{N}(Y_{2Ni}-\bar{Y}_{2N})^4\right\}^{1/2} \\ \leq& \frac{1}{N-1}\left\{\sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})^4 \vee \sum_{i=1}^{N}(Y_{2Ni}-\bar{Y}_{2N})^4\right\} \\ \leq& \frac{NL}{N-1}.\end{aligned}$$

So by (i), $\left|\sum_{i=1}^{N} W_{N,i}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})/N_1 - \sum_{i=1}^{N}(Y_{1Ni}-\bar{Y}_{1N})(Y_{2Ni}-\bar{Y}_{2N})/N\right| \xrightarrow{\text{as}} 0$.

In addition, $S_{12N} \leq \|S_N\|_F$ is bounded from above. These imply that

$$\begin{aligned}
|\tilde{S}_{12N} - S_{12N}| =& \Big|\frac{N_1}{N_1 - 1}\Big\{\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{1Ni} - \bar{Y}_{1N})(Y_{2Ni} - \bar{Y}_{2N}) \\
&- \frac{1}{N}\sum_{i=1}^{N}(Y_{1Ni} - \bar{Y}_{1N})(Y_{2Ni} - \bar{Y}_{2N})\Big\} \\
&+ \big(\frac{N_1}{(N_1 - 1)N} - \frac{1}{N - 1}\big)\sum_{i=1}^{N}(Y_{1Ni} - \bar{Y}_{1N})(Y_{2Ni} - \bar{Y}_{2N})\Big| \\
\leq& \frac{N_1}{N_1 - 1}\Big|\frac{1}{N_1}\sum_{i=1}^{N} W_{N,i}(Y_{1Ni} - \bar{Y}_{1N})(Y_{2Ni} - \bar{Y}_{2N}) \\
&- \frac{1}{N}\sum_{i=1}^{N}(Y_{1Ni} - \bar{Y}_{1N})(Y_{2Ni} - \bar{Y}_{2N})\Big| + \frac{N - N_1}{(N - 1)(N_1 - 1)}\frac{N - 1}{N}S_{12N} \to 0.
\end{aligned}$$

We now finally have $|\hat{S}_{12N} - S_{12N}| \leq |\hat{S}_{12N} - \tilde{S}_{12N}| + |\tilde{S}_{12N} - S_{12N}| \xrightarrow{\text{as}} 0$. □

If we use $Y_{2Ni} \leftarrow Y_{1Ni}$ above, then we get the same proof as item (ii) of Lemma 3.

Lemma 6. *Under Assumption D and for all sequences of W, the imputed potential outcomes satisfy $\lim_{N\to\infty} \max_{i,j} |Y_i^*(j) - \bar{Y}^*(j)|^2/N = 0$.*

Proof. From (5.3), we obtain $\{Y_i^*(j)_1 : i = 1, \ldots, N, j = 1, \ldots, J\}$ from $\{W_i, (Y_i^{\text{obs}})_1 : i = 1, \ldots, N\}$ in the same way as FRT-2. So by Lemma 4, we have $\lim_{N\to\infty} \max_{i,j}\{Y_i^*(j)_1 - \bar{Y}^*(j)_1\}^2/N = 0$. Doing the same for the other $d - 1$ entries gives the desired result. □

A.2 Proofs of Main Text Results

We make some preliminary observations and extend the notation to handle the permutation distributions as required by Theorems 1, 2, and 3. Throughout, we make heavy use of the mean of the observed values:

$$\bar{Y}_\bullet^{\text{obs}} = \frac{1}{N}\sum_{i=1}^{N} Y_i^{\text{obs}} = \sum_{j=1}^{J}\frac{N_j}{N}\hat{\bar{Y}}(j)$$

Recall the imputed potential outcomes FRT-2 are $Y_i^*(j) = Y_i^{\text{obs}} + z_j - z_{W_i}$. They agree with the data in the sense $Y_i^*(W_i) = Y_i^{\text{obs}}$ for all $i = 1, \ldots, N$. They are also strictly additive, as $Y_i^*(j) - Y_i^*(k) = z_j + Y_i^{\text{obs}} - z_{W_i} - (z_k + Y_i^{\text{obs}} - z_{W_i}) = z_j - z_k$ does not depend on the unit i. Define $\bar{z} = \sum_{j=1}^{J} N_j z_j/N$. Their means are $\bar{Y}^* = (\bar{Y}^*(1), \ldots, \bar{Y}^*(J))^\top$. Their covariance structure is $s^* 1_J 1_J^\top$, due to strict additivity. We have

$$\bar{Y}^*(j) = \frac{1}{N}\sum_{i=1}^{N}(Y_i^{\text{obs}} + z_j - z_{W_i}) = \sum_{k=1}^{J}\frac{N_k}{N}\hat{\bar{Y}}(k) + z_j - \bar{z}, \tag{A.1}$$

$$s^* = S^*(1,1) = \frac{1}{N-1}\sum_{i=1}^{N}\{Y_i^*(1) - \bar{Y}^*(1)\}^2 = \frac{1}{N-1}\sum_{j=1}^{J}\sum_{i=1}^{N} W_i(j)\{Y_i^*(j) - \bar{Y}^*(j)\}^2,$$

$$= \sum_{j=1}^{J}\frac{N_j - 1}{N-1}\hat{S}(j,j) + \sum_{j=1}^{J}\frac{N_j}{N-1}\{\hat{\bar{Y}}(j) - \bar{Y}^*(j)\}^2,$$

where the second to last equality is because $Y_i^*(j) - \bar{Y}^*(j)$ does not depend on j due to strict additivity and $\sum_{j=1}^{J} W_i(j) = 1$. The last equality follows from the bias-variance decomposition (add and subtract $\hat{\bar{Y}}(j)$) and noting $Y_i^*(j) = Y_i^{\text{obs}}$ when $W_i = j$. For asymptotic purposes, note that $C, x, \tilde{C}, \tilde{x}$ are fixed with respect to N, hence z is as well. They may be regarded as constants as we take $N \to \infty$.

The analogs of $\hat{D}$ and V, for imputed potential outcomes are, respectively

$$\hat{D}_\pi = N \cdot \text{diag}\{\hat{S}_\pi(1,1)/N_1, \ldots, \hat{S}_\pi(J,J)/N_J\}, \quad V^* = s^*(P^{-1} - 1_J 1_J^\top). \tag{A.2}$$

Compare these to (2.4) and (2.3). We also have, conditional on W, that $\hat{D}_\pi - s^* P^{-1} \xrightarrow{P} 0$. In general, consistent with previous patterns, analogs of population quantities have superscript "$*$", while those of observed quantities have subscript "π".

Proof of Theorems 1, 2, and 3. We prove the randomization, followed by the permutation distribution claims.

Randomization distributions of X^2, F, and B. Let Assumption A and $H_{0\text{N}}(C,x)$ hold. We have $N^{1/2}(C\hat{\bar{Y}} - x) \xrightarrow{\text{d}} \mathcal{N}(0_m, CVC^\top)$, $C\hat{D}C^\top \xrightarrow{P} CDC^\top \succ 0$ and $CDC^\top \succeq CVC^\top$ by Proposition 3 and (2.4). Hence, by Lemma 1

$$X^2 = N^{1/2}(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1} N^{1/2}(C\hat{\bar{Y}} - x) \xrightarrow{\text{d}} \sum_{j=1}^{m} a_j \xi_j^2, \quad \text{with } a_j \in [0,1] \quad (j = 1, \ldots, m).$$

We deal with B, F similarly. Assume $x = 0_m$. By (2.4) and the Continuous Mapping Theorem, $\text{tr}(M\hat{D})CC^\top \xrightarrow{P} \text{tr}(MD)CC^\top$, so

$$B = N^{1/2}(C\hat{\bar{Y}})^\top (\text{tr}(M\hat{D})CC^\top)^{-1} N^{1/2} C\hat{\bar{Y}} \xrightarrow{\text{d}} \sum_{j=1}^{m} \lambda_j \big(CVC^\top (\text{tr}(MD)CC^\top)^{-1}\big)\xi_j^2$$

$$\overset{\text{d}}{=} \sum_{j=1}^{m} \frac{1}{\text{tr}(MD)} \lambda_j (VC^\top (CC^\top)^{-1} C)\xi_j^2 \overset{\text{d}}{=} \frac{\sum_{j=1}^{m} \lambda_j (MV)\xi_j^2}{\text{tr}(MD)}.$$

where the "$\xrightarrow{\text{d}}$" is by Lemma 1, and the first "$\overset{\text{d}}{=}$" is by Lemma 8. Recall $\mathcal{X}$ and $\hat{\sigma}^2$ in (3.3). Then $\hat{\sigma}^2 \xrightarrow{P} \sum_{j=1}^{J} p_j S(j,j) = \bar{S}$ by Proposition 2, $(N_j - 1)/(N - J) \to p_j$, and

$$(\mathcal{X}^\top \mathcal{X}/N)^{-1} = \text{diag}(N_1/N, \ldots, N_J/N)^{-1} \xrightarrow{P} P^{-1}.$$

Therefore, by Lemma 1,

$$mF = N^{1/2}(C\hat{\bar{Y}})\{\hat{\sigma}^2 C(\mathcal{X}^\top \mathcal{X})^{-1} C^\top\}^{-1} N^{1/2} C\hat{\bar{Y}} \xrightarrow{\text{d}} \sum_{j=1}^{m} \lambda_j (CVC^\top (\bar{S}CP^{-1}C^\top)^{-1})\xi_j^2.$$

Permutation distributions. One way to proceed is to show the imputed potential outcomes satisfy Assumption A almost surely. By inspection of (A.1) and invoking Lemma 3, the means $\bar{Y}^*$ and covariances $s^* 1_J 1_J^\top$ of the imputed potential outcomes converge almost surely under Assumption B. However, we make a slightly more interesting argument here, and save the above approach for the SRE and vector potential outcomes cases.

We first show, for almost all realizations of the sequence of treatment assignments W, that Assumption A holds for $\{U_i^*(j) : i = 1, \ldots, N, j = 1, \ldots, J\}$ where $U_i^*(j) = \{Y_i^*(j) - \bar{Y}^*(j)\}/(s^*)^{1/2}$ are the standardized imputed potential outcomes. Clearly they always have mean 0 and variance 1, so it is enough to verify that, almost surely

$$\lim_{N\to\infty} \max_{i,j} \frac{1}{N}\{U_i^*(j) - \bar{U}^*(j)\}^2 = \lim_{N\to\infty} \max_{i,j} \frac{\{Y_i^*(j) - \bar{Y}^*(j)\}^2}{Ns^*} = 0. \tag{A.3}$$

Starting with (A.1), we have

$$\begin{aligned} s^* =& \sum_{j=1}^J \frac{N_j - 1}{N-1}\hat{S}(j,j) + \sum_{j=1}^J \frac{N_j}{N-1}\{\hat{\bar{Y}}(j) - \bar{Y}^*(j)\}^2 \\ \geq& \frac{N_1 - 1}{N-1}\hat{S}(1,1) \xrightarrow{\text{as}} p_1 S(1,1), \end{aligned}$$

where the last step is by Lemma 3. This shows the sequence $(s^*)_{N\geq 2J}$ is bounded away from 0, as $p_1, S(1,1) > 0$. Now we also have $\lim_{N\to\infty} N^{-1} \max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2 = 0$, no matter what the realization of the sequence $\{W\}_{N=1}^\infty$ is, by Lemma 4. These two facts together show (A.3).

Because $\hat{S}(1,1) \xrightarrow{\text{as}} S(1,1)$ by Lemma 3, we for the rest of the proof fix a sequence of W along which $\hat{S}(1,1) \to S(1,1)$. The only remaining randomness then comes from $\pi \sim \text{Unif}(\Pi_N)$. Note for $i = 1, \ldots, N$ that $CU_i^* = C(Y_i^* - \bar{Y}^*)/(s^*)^{1/2} = 0_m$ because $CY_i^* = x$ from the fact that the imputed potential outcomes satisfy (2.2). In particular, the standardized imputed potential outcomes satisfy $H_{0\text{N}}(C, 0_m)$, ie $C\bar{U}^* = 0_m$. Hence, by Proposition 3, we have

$$\begin{aligned} (N/s^*)^{1/2}(C\hat{\bar{Y}}_\pi - x) =& N^{1/2} C(\hat{\bar{Y}}_\pi - \bar{Y}^*)/(s^*)^{1/2} = N^{1/2} C\hat{\bar{U}}_\pi \\ \xrightarrow{\text{d}}& \mathcal{N}\big(0_m, C(P^{-1} - 1_J 1_J^\top)C^\top\big) \overset{\text{d}}{=} \mathcal{N}(0_m, CP^{-1}C^\top) \end{aligned}$$

because the standardized imputed potential outcomes have covariance structure $1_J 1_J^\top$ and $C1_J = 0_m$. Next, for $j = 1, \ldots, J$, we have

$$\frac{\hat{S}_\pi(j,j)}{s^*} = \frac{1}{N-1}\sum_{i=1}^N W_{\pi(i)}(j)\frac{\{Y_i^*(j) - \bar{Y}^*(j)\}^2}{s^*} = \frac{1}{N-1}\sum_{i=1}^N W_{\pi(i)}(j) U_i^*(j)^2 \xrightarrow{P} 1$$

by Proposition 2 and because the standardized imputed potential outcomes have group variances 1. It follows by (A.2) that

$$\hat{D}_\pi/s^* \xrightarrow{P} P^{-1}, \quad \hat{\sigma}_\pi^2/s^* = \sum_{j=1}^J \frac{N_j - 1}{(N-J)s^*}\hat{S}_\pi(j,j) \xrightarrow{P} 1, \quad \text{tr}(M\hat{D}_\pi)/s^* \xrightarrow{P} \text{tr}(MP^{-1}).$$

We thus finally have by Lemma 1

$$\begin{aligned} X_\pi^2 =& (N/s^*)^{1/2}(C\hat{\bar{Y}}_\pi - x)^\top (C\hat{D}_\pi C^\top/s^*)^{-1}(N/s^*)^{1/2}(C\hat{\bar{Y}}_\pi - x) \\ \xrightarrow{\mathrm{d}} & \sum_{j=1}^m \lambda_j \big(CP^{-1}C^\top (CP^{-1}C^\top)^{-1}\big)\xi_j^2 \overset{\mathrm{d}}{=} \chi_m^2, \end{aligned}$$

and with $x = 0_m$ for the B and F statistics:

$$\begin{aligned} B_\pi =& (N/s^*)^{1/2}(C\hat{\bar{Y}}_\pi)^\top \{\mathrm{tr}(M\hat{D}_\pi)CC^\top/s^*\}^{-1}(N/s^*)^{1/2}C\hat{\bar{Y}}_\pi \\ \xrightarrow{\mathrm{d}} & \sum_{j=1}^m \lambda_j \big(CP^{-1}C^\top (\mathrm{tr}(MP^{-1})CC^\top)^{-1}\big)\xi_j^2 \overset{\mathrm{d}}{=} \sum_{j=1}^m \lambda_j(MP^{-1})\xi_j^2/\,\mathrm{tr}(MP^{-1}), \\ mF_\pi =& (N/s^*)^{1/2}(C\hat{\bar{Y}}_\pi)^\top \left\{ \frac{\hat{\sigma}_\pi^2}{s^*} C(\mathcal{X}^\top\mathcal{X}/N)^{-1}C^\top \right\}^{-1} (N/s^*)^{1/2}C\hat{\bar{Y}}_\pi \\ \xrightarrow{\mathrm{d}} & \sum_{j=1}^m \lambda_j \big(CP^{-1}C^\top (CP^{-1}C^\top)^{-1}\big)\xi_j^2 \overset{\mathrm{d}}{=} \chi_m^2, \end{aligned}$$

where the "$\xrightarrow{\mathrm{d}}$" for B_π uses $\lambda_j(CP^{-1}C^+) = \lambda_j(MP^{-1})$ by Lemma 8. □

From $s^* \geq p_1 S(1,1)$ a.s., we see that we do not need the full strength of Assumption B. It is enough that $\{Y_i(j) : i = 1, \ldots, N\}$ has finite fourth moment for some j.

Extending Theorem 1 to the case of stratified experiments or vector potential outcomes is straightforward. We also supply their proofs for completeness.

Proof of Theorem 4. We prove the randomization, followed by the permutation distribution claims.

Randomization distribution of X^2. For $h = 1, \ldots, H$, we have that $\mathbb{E}\hat{\bar{Y}}_{[h]} = \bar{Y}_{[h]}$, and that Assumption A holds in each stratum h. By Proposition 3,

$$N_{[h]}^{1/2} C(\hat{\bar{Y}}_{[h]} - \bar{Y}_{[h]}) \xrightarrow{\mathrm{d}} \mathcal{N}(0_m, CV_{[h]}C^\top), \text{ where } V_{[h]} = \operatorname*{plim}_{N\to\infty} \hat{D}_{[h]} - S_{[h]}.$$

Under $H_{0\mathrm{N}}(C, x)$, we have $x = C\bar{Y} = \sum_{h=1}^H N_{[h]} C\bar{Y}_{[h]}/N$. Because $(\hat{\bar{Y}}_{[1]}, \ldots, \hat{\bar{Y}}_{[H]})$ are mutually independent in a SRE, we have

$$\begin{aligned} N^{1/2}(C\breve{\bar{Y}} - x) =& \sum_{h=1}^H \left(\frac{N_{[h]}}{N}\right)^{1/2} N_{[h]}^{1/2} C(\hat{\bar{Y}}_{[h]} - \bar{Y}_{[h]}) \\ & \xrightarrow{\mathrm{d}} \sum_{h=1}^H \omega_{[h]}^{1/2} \mathcal{N}(0_m, CV_{[h]}C^\top) \overset{\mathrm{d}}{=} \mathcal{N}\left(0_m, \sum_{h=1}^H \omega_{[h]} CV_{[h]}C^\top\right). \end{aligned}$$

Next, note that $\operatorname{plim}_{N\to\infty} \hat{D}_{[h]} \succeq V_{[h]}$ implies

$$\operatorname*{plim}_{N\to\infty} \sum_{h=1}^{H} N_{[h]} C \hat{D}_{[h]} C^\top / N \succeq \sum_{h=1}^{H} \omega_{[h]} C V_{[h]} C^\top,$$

so by Lemma 1, we have

$$X^2 = N^{1/2}(C\breve{Y} - x)^\top \left(C \sum_{h=1}^{H} \frac{N_{[h]}}{N} \hat{D}_{[h]} C^\top \right)^{-1} N^{1/2}(C\breve{Y} - x) \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2.$$

Permutation distribution of X^2. We first show Assumption A holds almost surely within each stratum for the imputed potential outcomes $Y_i^*(j)$. Because the original potential outcomes satisfy Assumption A in each stratum, Lemma 4 gives

$$\lim_{N\to\infty} \max_j \max_{i:X_i=h} \{Y_i^*(j) - \bar{Y}_{[h]}^*(j)\}^2 / N_{[h]} = 0.$$

Put $\bar{z}_{[h]} = \sum_{j=1}^{J} N_{[h]j} z_{[h],j} / N_{[h]}$. In stratum h, the mean vector is $\bar{Y}_{[h]}^*$ and the covariance structure is $s_{[h]}^* 1_J 1_J^\top$, where

$$\begin{aligned} \bar{Y}_{[h]}^*(j) &= \sum_{k=1}^{J} \frac{N_{[h]k}}{N_{[h]}} \hat{\bar{Y}}_{[h]}(k) + z_{[h],j} - \bar{z}_{[h]} \\ s_{[h]}^* &= \sum_{j=1}^{J} \frac{N_{[h]j} - 1}{N_{[h]} - 1} \hat{S}_{[h]}(j,j) + \sum_{j=1}^{J} \frac{N_{[h]j}}{N_{[h]} - 1} \{\hat{\bar{Y}}_{[h]}(j) - \bar{Y}_{[h]}^*(j)\}^2, \end{aligned}$$

by applying (A.1) to stratum h. $\bar{Y}_{[h]}^*$ and $s_{[h]}^*$ converge almost surely because all quantities on the right-hand side do. $\hat{\bar{Y}}_{[h]}(j)$ and $\hat{S}_{[h]}(j,j)$ converge almost surely because of Lemma 3, applicable because Assumption B holds within stratum h. This shows Assumption A holds within each stratum almost surely.

For the rest of the proof, fix a sequence (W) along which $(s_{[h]}^*)$ converges. Because each $CY_i^* = x_{[h]}$ whenever $X_i = h$, we have $C\bar{Y}_{[h]}^* = x_{[h]}$, and by Proposition 3,

$$N_{[h]}^{1/2} C(\hat{\bar{Y}}_{[h],\pi} - \bar{Y}_{[h]}^*) \xrightarrow{\mathrm{d}} \mathcal{N}\big(0_m, s_{[h]}^* C(P^{-1} - 1_J 1_J^\top) C^\top\big) \overset{\mathrm{d}}{=} \mathcal{N}(0_m, s_{[h]}^* C P^{-1} C^\top).$$

Since $x = \sum_{h=1}^{H} N_{[h]} x_{[h]} / N = \sum_{h=1}^{H} N_{[h]} C \bar{Y}_{[h]}^* / N$, it follows that

$$\begin{aligned} N^{1/2}(C\breve{\bar{Y}}_\pi - x) &= \sum_{h=1}^{H} \left(\frac{N_{[h]}}{N} \right)^{1/2} N_{[h]}^{1/2} C(\hat{\bar{Y}}_{[h],\pi} - \bar{Y}_{[h]}^*) \\ &\xrightarrow{\mathrm{d}} \sum_{h=1}^{H} \omega_{[h]}^{1/2} \mathcal{N}\left(0_m, s_{[h]}^* C P^{-1} C^\top\right) \overset{\mathrm{d}}{=} \mathcal{N}\left(0_m, \sum_{h=1}^{H} \omega_{[h]} s_{[h]}^* C P^{-1} C^\top\right) \end{aligned}$$

because, conditioning on W, the $(\hat{\breve{Y}}_{[1],\pi}, \ldots, \hat{\breve{Y}}_{[H],\pi})$ are mutually independent. Next, from Proposition 2, we have $\hat{D}_{[h],\pi} \xrightarrow{P} s^*_{[h]} P^{-1}$, so $C \sum_{h=1}^{H} N_{[h]} \hat{D}_{[h],\pi} C^\top / N \xrightarrow{P} \sum_{h=1}^{H} \omega_{[h]} s^*_{[h]} C P^{-1} C^\top$, and we finally have from Lemma 1

$$X^2_\pi = N^{1/2}(C\breve{\bar{Y}}_\pi - x)^\top \left(C \sum_{h=1}^{H} \frac{N_{[h]}}{N} \hat{D}_{[h],\pi} C^\top \right)^{-1} N^{1/2}(C\breve{\bar{Y}}_\pi - x) \xrightarrow{\mathrm{d}} \chi^2_m. \qquad \square$$

The main idea of the proof is to apply the same concepts of the proof of the CRE case in Theorem 1 stratum by stratum. A key difference is that, in the CRE case, we could standardize all the imputed potential outcomes by a scalar s^*. Even though, in the SRE case, we can standardize by $s^*_{[h]}$ strata, there is no obvious single quantity by which to standardize all the imputed potential outcomes. We thus had to show that each $s^*_{[h]}$ converged, rather than just being bounded away from zero.

Proof of Theorem 5. We prove the randomization, followed by the permutation distribution claims.

Randomization distribution of X^2. This part has identical logic to proving Theorem 1. Let Assumption D and $H_{0\mathrm{N}}(C, x)$ hold. We have $N^{1/2}(C\hat{\bar{Y}} - x) \xrightarrow{\mathrm{d}} \mathcal{N}(0_m, CVC^\top)$, $C\hat{D}C^\top \xrightarrow{P} CDC^\top \succ 0$ and $CDC^\top \succeq CVC^\top$ by Proposition 8. Hence, by Lemma 1

$$X^2 = N^{1/2}(C\hat{\bar{Y}} - x)^\top (C\hat{D}C^\top)^{-1} N^{1/2}(C\hat{\bar{Y}} - x) \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j \xi_j^2,$$

with $a_j \in [0, 1]$, $j = 1, \ldots, m$.

Permutation distribution of X^2. We first show Assumption D holds almost surely for the imputed potential outcomes $Y^*_i(j)$. Because the original potential outcomes satisfy Assumption D, Lemma 6 gives $\lim_{N\to\infty} \max_{i,j} |Y^*_i(j) - \bar{Y}^*(j)|^2 / N = 0$. Their means satisfy

$$\bar{Y}^*(j)_1 = \frac{1}{N} \sum_{i=1}^{N} (z_{1j} + Y^{\mathrm{obs}}_{i,1} - z_{1,W_i}) = z_{1j} + \frac{1}{N} \sum_{j=1}^{J} N_j \hat{\bar{Y}}(j)_1 - \bar{z}_1.$$

Hence, the $\bar{Y}^*(j)_1$ converge almost surely because $\hat{\bar{Y}}(j) \xrightarrow{\mathrm{as}} \bar{Y}(j)$ by Lemma 5. By the same reasoning, the other entries of $\bar{Y}^*(j)$ also converge almost surely. The covariance structure of the imputed potential outcomes is $(1_J 1_J^\top) \otimes S^*(1, 1)$, where following the same steps to

derive (A.1), we get

$$\begin{aligned}
S^*(1,1) =& \frac{1}{N-1}\sum_{i=1}^{N}\{Y_i^*(1)-\bar{Y}^*(1)\}\{Y_i^*(1)-\bar{Y}^*(1)\}^\top \\
=& \frac{1}{N-1}\sum_{j=1}^{J}\sum_{i=1}^{N} W_i(j)\{Y_i^*(j)-\bar{Y}^*(j)\}\{Y_i^*(j)-\bar{Y}^*(j)\}^\top \\
=& \sum_{j=1}^{J}\frac{N_j-1}{N-1}\hat{S}(j,j) + \sum_{j=1}^{J}\frac{N_j}{N-1}\{\hat{\bar{Y}}(j)-\bar{Y}^*(j)\}\{\hat{\bar{Y}}(j)-\bar{Y}^*(j)\}^\top.
\end{aligned}$$

This converges almost surely because all quantities in the last line do. For instance, $\hat{S}(j,j)$ converge almost surely because of Lemma 5, applicable because of Assumption E. This shows Assumption D holds almost surely.

For the rest of the proof, fix a sequence W along which Assumption D is met. The limit of $S^*(1,1)$ must be invertible because the above calculation shows $S^*(1,1) \succeq (N_1-1)S(1,1)/(N-1) \succ 0$. Because each $CY_i^* = x$, Proposition 8 gives us

$$\begin{aligned}
N^{1/2}(C\hat{\bar{Y}}_\pi - x) = N^{1/2}C(\hat{\bar{Y}}_\pi - \bar{Y}^*) \xrightarrow{\mathrm{d}} & \mathcal{N}\big(0_m, C\{(P^{-1}-1_J1_J^\top)\otimes S^*(1,1)\}C^\top\big) \\
\overset{\mathrm{d}}{=} & \mathcal{N}\big(0_m, C\{P^{-1}\otimes S^*(1,1)\}C^\top\big).
\end{aligned}$$

The cancellation in the last line occurred, for instance because the $(1,2)$-block of $C\{(1_J1_J^\top)\otimes S^*(1,1)\}C^\top$ is $(C_1\otimes e_1^\top)\{(1_J1_J^\top)\otimes S^*(1,1)\}(C_2\otimes e_2^\top)^\top = (C_1 1_J 1_J^\top C_2^\top)\otimes\{e_1^\top S^*(1,1)e_2\}$, which vanishes because C_1, C_2 are themselves contrast matrices. Next,

$$\hat{D}_\pi \xrightarrow{P} \operatorname{diag}\left\{\frac{S^*(1,1)}{p_1},\ldots,\frac{S^*(1,1)}{p_J}\right\} = P^{-1}\otimes S^*(1,1),$$

so $C\hat{D}_\pi C^\top \xrightarrow{P} C\{P^{-1}\otimes S^*(1,1)\}C^\top$, and we finally have from Lemma 1 that $X_\pi^2 = N(C\hat{\bar{Y}}_\pi - x)^\top(C\hat{D}_\pi C^\top)^{-1}(C\hat{\bar{Y}}_\pi - x) \xrightarrow{\mathrm{d}} \chi_m^2$. □

It is also difficult to use the standardized imputed potential outcomes $S^*(1,1)^{-1/2}\{Y_i^*(j)-\bar{Y}^*(j)\}$, even though they could help bypass having to show $S^*(1,1)$ converged. It is not entirely obvious how to express X_π^2 in terms of these standardized quantities. While $\operatorname{diag}(A,B) = C\cdot\operatorname{diag}(A/C, B/C)$ when $C\neq 0$ is a scalar, there is no analog if C is itself a matrix.

Proofs of other results in the main text.

Proof of Proposition 1. The conclusion follows from

$$\begin{aligned}
\max_{i,j}\frac{1}{N}\{Y_i(j)-\bar{Y}(j)\}^2 =& \frac{1}{N}\left[\max_{i,j}\{Y_i(j)-\bar{Y}(j)\}^4\right]^{1/2} \\
\leq& \frac{1}{N}\left[\max_{j}\sum_{i=1}^{N}\{Y_i(j)-\bar{Y}(j)\}^4\right] \leq (L/N)^{1/2}
\end{aligned}$$

which $\to 0$ as $N \to \infty$. □

Our next goal is to prove Proposition 4. Despite its statement being quite intuitive, we find the actual proof not to be all that trivial.

Proof of Proposition 4. This proof relies on Lemma 11. Assume $H_{0\mathrm{N}}(C, x)$ throughout. Define

$$F(x) := \mathbb{P}(T \le x),\ G(x) := \mathbb{P}(T < x),\ F_W(x) := \mathbb{P}(T_\pi \le x|W),\ G_W(x) := \mathbb{P}(T_\pi < x|W).$$

Let $U \sim \mathrm{Unif}(0, 1)$. We will show that

$$\mathbb{P}\{\frac{1}{N!}\sum_{\pi\in\Pi_N} 1(T_\pi \ge T) \le \alpha\} \le \alpha \text{ for all } \alpha \in (0,1) \iff T \le_{\mathrm{st}} T_\pi|W.$$

By definition, the FRT with test statistic T successfully controls type I error at all levels α when the left hand side is true.

We first show the sufficiency of $T \le_{\mathrm{st}} T_\pi|W$. Fix $\alpha \in (0, 1)$. Note that $G_W(T) = (N!)^{-1}\sum_{\pi\in\Pi_N} 1(T_\pi < T)$, so

$$\mathbb{P}\big\{\frac{1}{N!}\sum_{\pi\in\Pi_N} 1(T_\pi \ge T) \le \alpha\big\} = \mathbb{P}\{1 - G_W(T) \le \alpha\} \le \mathbb{P}\{G(T) \ge 1 - \alpha\} \le \mathbb{P}(U \ge 1 - \alpha) = \alpha$$

where we have used $T \le_{\mathrm{st}} T_\pi|W$ if and only if $G_W \le G$ on $\mathbb{R}$ and $G(T) \le_{\mathrm{st}} U$.

Now we show the necessity of $T \le_{\mathrm{st}} T_\pi|W$. If for some W it is not true that $T \le_{\mathrm{st}} T_\pi|W$, then there exists $x \in \mathbb{R}$ such that $F(x) < G_W(x)$ (this is because $F = G$, $F_W = G_W$, Lebesgue almost everywhere), pick $\alpha \in (1 - F(x), 1 - G_W(x))$. Then we fail to control type I error because

$$\begin{aligned}\mathbb{P}\big\{\frac{1}{N!}\sum_{\pi\in\Pi_N} 1(T_\pi \ge T) \le \alpha\big\} \ge& 1 - \mathbb{P}\{G_W(T) \le 1 - \alpha\} = 1 - F\big(\sup\{t : G_W(t) \le 1 - \alpha\}\big)\\ \ge& 1 - F(x) > \alpha\end{aligned}$$

where the second equality follows because $\{t : G_W(t) \le 1 - \alpha\}$ is closed (due to the left continuity of G_W). Its measure under the distribution of T is hence F evaluated at its right endpoint. The second $\ge$ follows because $G_W(t) \le 1 - \alpha < G_W(x)$ implies $t \le x$ (as G_W is nondecreasing), so $\sup\{t : G_W(t) \le 1 - \alpha\} \le x$. □

Proof of Corollary 1. First, if $S(1,1) = \cdots = S(J,J)$, then $D = S(1,1)P^{-1}$ from (2.4). Recall from $V \preceq D$ that each

$$\lambda_j(MV) = \lambda_j(M^2V) = \lambda_j(MVM) \le \lambda_j(MDM) = \lambda_j(MD) \tag{A.4}$$

Therefore, under $H_{0\mathrm{N}}(C, x)$, Theorem 2 implies that

$$\begin{aligned}B \xrightarrow{\mathrm{d}} \frac{\sum_{j=1}^m \lambda_j(MV)\xi_j^2}{\mathrm{tr}(MD)} \le_{\mathrm{st}} \frac{\sum_{j=1}^m \lambda_j(MD)\xi_j^2}{\mathrm{tr}(MD)} =& \frac{\sum_{j=1}^m S(1,1)\lambda_j(MP^{-1})\xi_j^2}{S(1,1)\,\mathrm{tr}(MP^{-1})}\\ =& \frac{\sum_{j=1}^m \lambda_j(MP^{-1})\xi_j^2}{\mathrm{tr}(MP^{-1})} \overset{\mathrm{d}}{=} B_\pi|W.\end{aligned}$$

So the criterion of Proposition 4 is met.

Second, if C is a row vector, then $M = C^\top C/CC^\top$. Therefore

$$B = \frac{\hat{\tilde{Y}}^\top C^\top C\hat{\tilde{Y}}/CC^\top}{\mathrm{tr}(C^\top C\hat{D})/CC^\top} = \frac{(C\hat{\tilde{Y}})^\top C\hat{\tilde{Y}}}{C\hat{D}C^\top} = (C\hat{\tilde{Y}})^\top (C\hat{D}C^\top)^{-1} C\hat{\tilde{Y}} = X^2. \qquad \square$$

Proof of Proposition 5. Under a balanced design we have $N_1 = \ldots = N_J = N/J$, $X^\top X = N_1 I_J$ and $\hat{\sigma}^2 = \sum_{j=1}^J \hat{S}(j,j)/J$. Thus, $F = N_1 \hat{\tilde{Y}}^\top M\hat{\tilde{Y}}/(m\hat{\sigma}^2)$. If M has the same values on its main diagonal, then each value is in fact m/J because the trace and rank of a projection matrix are the same. This implies

$$\begin{aligned} \frac{N}{\mathrm{tr}(M\hat{D})} =& N/\left\{\sum_{j=1}^J \frac{N}{N_j}\hat{S}(j,j)\frac{m}{J}\right\} = \frac{N}{m\sum_{j=1}^J \hat{S}(j,j)} = \frac{N_1}{m\hat{\sigma}^2} \\ \Longrightarrow B =& \frac{N(\hat{\tilde{Y}})^\top M\hat{\tilde{Y}}}{\mathrm{tr}(M\hat{D})} = \frac{N_1\hat{\tilde{Y}}^\top M\hat{\tilde{Y}}}{m\hat{\sigma}^2} = F. \end{aligned} \qquad \square$$

Proof of Corollary 2. If $S(1,1) = \cdots = S(J,J)$, then $\bar{S} = \sum_{j=1}^J p_j S(j,j) = S(1,1)$ and $D = \bar{S}\cdot P^{-1}$. Therefore, $0 \le \lambda_j\big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big) = \lambda_j\big(CVC^\top(CDC^\top)^{-1}\big) \le 1$ because $V \preceq D$. By Theorem 3, under $H_{0\mathrm{N}}(C, 0_m)$, we have

$$m\cdot F \xrightarrow{\mathrm{d}} \sum_{j=1}^m \lambda_j\big(CVC^\top(\bar{S}CP^{-1}C^\top)^{-1}\big)\xi_j^2 \le_{\mathrm{st}} \chi^2_m, \qquad m\cdot F_\pi | W \xrightarrow{\mathrm{d}} \chi^2_m. \qquad \square$$

Proof of Proposition 6. The conclusions follow from simple linear algebra facts. They seem to be known, but we give a proof for completeness.

We first equate the X^2. As stated, in the ANOVA setting, $C = (1_{J-1},\ -I_{J-1})$ and $x = 0_{J-1}$. Put $Q_j = N_j/\hat{S}(j,j)$ and $Q = \sum_{j=1}^J Q_j$. Then by block matrix multiplication

$$\begin{aligned} \frac{1}{N}C\hat{D}C^\top =& (1_{J-1},\ -I_{J-1})\,\mathrm{diag}(1/Q_1,\ldots,1/Q_J)\begin{pmatrix} 1_{J-1}^\top \\ -I_{J-1}\end{pmatrix} \\ =& \frac{1}{Q_1}1_{J-1}1_{J-1}^\top + \mathrm{diag}(1/Q_2,\ldots,1/Q_J). \end{aligned}$$

Thus, using the Sherman–Morrison formula $(A+uv^\top)^{-1} = A^{-1} - A^{-1}uv^\top A^{-1}/(1+v^\top A^{-1}u)$, we have

$$\begin{aligned} \left(\frac{1}{N}C\hat{D}C^\top\right)^{-1} =& \mathrm{diag}(Q_2,\ldots,Q_J) - \left\{\frac{1}{Q_1}\begin{pmatrix} Q_2 \\ \vdots \\ Q_J\end{pmatrix}(Q_2,\ldots,Q_J)\right\} \Big/ \left\{1 + \frac{1}{Q_1}\sum_{j=2}^J Q_j\right\} \\ =& \mathrm{diag}(Q_2,\ldots,Q_J) - \frac{1}{Q}\begin{pmatrix} Q_2 \\ \vdots \\ Q_J\end{pmatrix}(Q_2,\ldots,Q_J). \end{aligned}$$

Finally, from (3.1), we have

$$X^2 = \begin{pmatrix} \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2) \\ \vdots \\ \hat{\bar{Y}}(1) - \hat{\bar{Y}}(J) \end{pmatrix} \left\{ \mathrm{diag}(Q_2, \ldots, Q_J) - \frac{1}{Q} \begin{pmatrix} Q_2 \\ \vdots \\ Q_J \end{pmatrix} (Q_2, \ldots, Q_J) \right\} \begin{pmatrix} \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2) \\ \vdots \\ \hat{\bar{Y}}(1) - \hat{\bar{Y}}(J) \end{pmatrix}$$
$$= \sum_{j=2}^{J} Q_j \{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(j)\}^2 - \frac{1}{Q} \left[\sum_{j=2}^{J} Q_j \{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(j)\} \right]^2 .$$

Now we recognize the expression in (4.2) as Q times the variance of $\{\hat{\bar{Y}}(1), \ldots, \hat{\bar{Y}}(J)\}$ under the probabilities $Q_1/Q, \ldots, Q_J/Q$. But variance is unaffected by switching signs, and then adding the constant $\hat{\bar{Y}}(1)$ to all quantities, so (4.2) is Q times the variance of $\{0, \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2), \ldots, \hat{\bar{Y}}(1) - \hat{\bar{Y}}(J)\}$ under the same probabilities, which is precisely what X^2 is above.

Next, we equate the F. Recall that $m = J - 1$. It is thus enough to show

$$(C\hat{\bar{Y}})^\top \{C(\mathcal{X}^\top \mathcal{X})^{-1} C^\top\}^{-1} C\hat{\bar{Y}} = \sum_{j=1}^{J} N_j \{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\mathrm{obs}}\}^2.$$

This follows an identical argument to showing the X^2 coincide, with N_j, N in place of Q_j, Q. □

Proof of Corollary 5. Under Assumption A and $H_{0\mathrm{N}}(C, x)$ with a row vector C, we have $N^{1/2}(C\hat{\bar{Y}} - x) \xrightarrow{\mathrm{d}} \mathcal{N}(0, CVC^\top)$ by Proposition 3, $C\hat{D}C^\top \xrightarrow{P} CDC^\top > 0$ and $CDC^\top \geq CVC^\top$ by (2.4). Hence,

$$t = \frac{N^{1/2}(x - C\hat{\bar{Y}})}{(C\hat{D}C^\top)^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, a), \text{ where } a = \frac{CVC^\top}{CDC^\top} \in [0, 1].$$

To show the permutation distribution under Assumption B, we have $\hat{S}(1,1) \xrightarrow{\mathrm{as}} S(1,1)$ by Lemma 3, so fix a sequence of W along which $\hat{S}(1,1) \to S(1,1)$. Then $(N/s^*)^{1/2}(C\hat{\bar{Y}}_\pi - x) \xrightarrow{\mathrm{d}} \mathcal{N}(0, CP^{-1}C^\top)$ and $\hat{D}_\pi/s^* \xrightarrow{P} P^{-1}$ (these are intermediate steps in the proof of Theorem 1), so

$$t_\pi | W = \frac{N^{1/2}(x - C\hat{\bar{Y}}_\pi)}{(C\hat{D}_\pi C^\top)^{1/2}} = (N/s^*)^{1/2} \frac{x - C\hat{\bar{Y}}_\pi}{(C\hat{D}_\pi C^\top)^{1/2}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

It thus holds that t_+ is proper for $H_{0\mathrm{N}}(C, x)$ by Proposition 4 and the fact that $\mathcal{N}(0, a)_+ \leq_{\mathrm{st}} \mathcal{N}(0, 1)_+$.

To argue t_+ is proper for (4.4), we let $x = x_0$. Then we want to test $\tilde{H}_{0\mathrm{N}}(C, x_0) : C\bar{Y} \geq x_0$. The notation switch frees up x as a dummy variable. Let $p(x)$ be the p-value from testing $C\bar{Y} = x$ with $t_+ = t_+(x)$. Then the p-value for $\tilde{H}_{0\mathrm{N}}(C, x)$ is $\sup_{x \geq x_0} p(x)$. When $x \leq C\hat{\bar{Y}}$, we have $t_+ = 0$, so $p(x) = 1$. If $C\hat{\bar{Y}} \geq x_0$, then $t_+(x_0) = 0$, so $p(x_0) = 1$ (see also the Hodges–Lehmann discussion), and $\sup_{x \geq x_0} p(x) = 1 = p(x_0)$. The more interesting case is

$C\hat{\tilde{Y}} < x_0$. Then $t_+(x_0) \leq t_+(x)$ when $x \geq x_0$. The fact that $t_\pi(x)|W \xrightarrow{\mathrm{d}} \mathcal{N}(0,1)$ a.s. for all $x \in \mathbb{R}$ suggests asymptotically that $p(x_0) \geq p(x)$ when $x \geq x_0$, so $\sup_{x \geq x_0} p(x) = p(x_0)$. Asymptotically speaking, we thus always have $\sup_{x \geq x_0} p(x) = p(x_0)$. This is why we can test $\tilde{H}_{0\mathrm{N}}(C, x_0)$ with t_+ as if we were testing $H_{0\mathrm{N}}(C, x)$. □

Appendix B

Extra background material

B.1 Proofs of other results

Here we state some additional results that could illuminate further the proofs in the first appendix. We start with some standard results on the convergence of random variables. More details can be found in [51].

Lemma 7. *Let $(Y_n), (A_n), (B_n)$ be sequences of random variables, Y a random variable, $a, b \in \mathbb{R}$. Then*

- *(*Continuous mapping theorem*) If f is continuous at a and $Y_n \xrightarrow{P} a$, then $f(Y_n) \xrightarrow{P} f(a)$*
- *If f is continuous and $Y_n \xrightarrow{\mathrm{d}} Y$, then $f(Y_n) \xrightarrow{\mathrm{d}} f(Y)$*
- *(*Slutsky's theorem*) If $Y_n \xrightarrow{\mathrm{d}} Y$, $A_n \xrightarrow{P} a$, $B_n \xrightarrow{P} b$, then $A_n + B_n Y_n \xrightarrow{\mathrm{d}} a + bY$*

We now collect some standard results from Linear Algebra. Recall that the *Loewner order* for positive semidefinite matrices $A, B \succeq 0$ is $A \succeq B$, which means $A - B \succeq 0$.

Lemma 8. *If $A \succeq B \succeq 0$ then*

- *If $B \succ 0$, then $A^{1/2} \succeq B^{1/2}$ and $B^{-1} \succeq A^{-1}$.*
- *We need not have $A^2 \succeq B^2$, unless $AB = BA$, in which case in fact $A^k \succeq B^k$ for all $k \in \mathbb{Z}^+$.*
- *If C has appropriate dimension, then $C^\top AC \succeq C^\top BC$.*
- *Each $\lambda_j(A) \geq \lambda_j(B)$ (in fact we only need for A, B to be symmetric).*
- *If $A, B^\top \in \mathbb{R}^{m \times n}$ and $m \geq n$, then $\det(AB - \lambda I_m) = (-\lambda)^{m-n} \det(BA - \lambda I_n)$. In particular, AB and BA have the same nonzero eigenvalues.*
- *If $A \succ 0$, then all $\lambda_j(BA^{-1}) \in [0, 1]$.*

Proof. All these results may be found, for instance, in [39]. However, we share an elegant proof of the fifth item [90]. Put

$$C = \begin{pmatrix} \lambda I_m & A \\ B & I_n \end{pmatrix},\ D = \begin{pmatrix} I_m & 0_{m\times n} \\ -B & \lambda I_n \end{pmatrix}$$

and use $\det(CD) = \det(DC)$.

For the sixth, note $\lambda_j(BA^{-1}) = \lambda_j(A^{-1/2}BA^{-1/2}) \leq \lambda_j(A^{-1/2}AA^{-1/2}) = 0$ by items 5 and 3, and $\lambda_j(A^{-1/2}BA^{-1/2}) \geq 0$. □

Survey sampling is the backbone of CREs. We now summarize some of its key results. Given data vectors $x, y \in \mathbb{R}^N$, which have mans $\bar{x} = \sum_{i=1}^N x_i/N$, $\bar{y}$, variances $S_x^2 = \sum_{i=1}^N (x_i - \bar{x})^2/(N-1)$, S_y^2, and covariance $S_{xy} = \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})/(N-1)$. Note if we put $V_N = I_N - 1_N 1_N^\top/N$, which is the projection matrix onto $\mathcal{C}(1_N)^\perp$, a handy formula for the (co)variance is then

$$S_x^2 = \frac{x^\top V_N x}{N-1},\ S_{xy} = \frac{x^\top V_N y}{N-1} \tag{B.1}$$

Recall $W \in \{1,\ldots,J\}^N$ is generated according to a CRE. This means each $\{i : W_i = j\}$ is a SRS of size N_j from $\{1,\ldots,N\}$. For $j = 1,\ldots,J$, put $\hat{\bar{x}}(j) = \sum_{i=1}^N W_i(j)x_i/N_j$ and $\hat{S}_x^2(j) = \sum_{i=1}^N W_i(j)\{x_i - \hat{\bar{x}}(j)\}^2/(N_j - 1)$. Define $\hat{\bar{y}}(j)$ and $\hat{S}_y^2(j)$ similarly, and $\hat{S}_{xy}(j) = \sum_{i=1}^N W_i(j)\{x_i - \hat{\bar{x}}(j)\}\{y_i - \hat{\bar{y}}(j)\}/(N_j - 1)$. This setup is helpful for inference on the mean and variance of a population based on the SRS counterparts.

Lemma 9.

- *Each* $\mathbb{E}W_i(j) = N_j/N$, *and*

$$\mathrm{Cov}\begin{pmatrix} W_1(j) \\ \vdots \\ W_N(j) \end{pmatrix} = \frac{N_j(N-N_j)}{N(N-1)} V_N,\ \mathrm{Cov}\left(\begin{pmatrix} W_1(j) \\ \vdots \\ W_N(j) \end{pmatrix}, \begin{pmatrix} W_1(k) \\ \vdots \\ W_N(k) \end{pmatrix}\right) = \frac{-N_j N_k}{N(N-1)} V_N$$

  when $j \neq k$. *In particular,* $\mathrm{Var}(W_1(j)) = N_j(N-N_j)/N^2$, $\mathrm{Cov}(W_1(j), W_2(j)) = N_j(N-N_j)/\{N^2(N-1)\}$, $\mathrm{Cov}(W_1(j), W_1(k)) = -N_jN_k/N^2$, *and* $\mathrm{Cov}(W_1(j), W_2(k)) = N_jN_k/\{N^2(N-1)\}$.

- *Each* $\mathbb{E}\hat{\bar{x}}(j) = \bar{x}$, *and*

$$\mathrm{Var}(\hat{\bar{x}}(j)) = \frac{N-N_j}{N \cdot N_j} S_x^2,\ \mathrm{Cov}(\hat{\bar{x}}(j), \hat{\bar{y}}(j)) = \frac{N-N_j}{N \cdot N_j} S_{xy},$$
$$\mathrm{Cov}(\hat{\bar{x}}(j), \hat{\bar{x}}(k)) = -S_x^2/N,\ \mathrm{Cov}(\hat{\bar{x}}(j), \hat{\bar{y}}(k)) = -S_{xy}/N$$

- *Each* $\mathbb{E}\hat{S}_x^2(j) = S_x^2$ *and* $\mathbb{E}\hat{S}_{xy} = S_{xy}$.

For the second item, note if $\hat{\bar{x}}(j)$ were instead the mean of N_j iid samples with replacement from the vector x, then $\mathrm{Var}(\hat{\bar{x}}(j)) = S_x^2/N_j$. If the samples are without replacement as in a SRS, then we have an additional multiplier $1 - N_j/N$ called the *finite population correction*. For the third item, the fact that $\hat{S}_x^2(j)$ is unbiased for S_x^2 motivates having the division by $N-1$ when we define population (co)variances S_x^2, S_y^2, and S_{xy}.

Proof. A common theme throughout the proof is that, though the $W_i(j)$'s are random, they sum over i to something deterministic. We also continually exploit the symmetry property $(W_{i(1)}(j), W_{i(2)}(k)) \stackrel{\mathrm{d}}{=} (W_1(j), W_2(k))$ for all $i_1 \neq i_2 \in \{1, \ldots, n\}$, and $j, k = 1, \ldots, J$.

Part 1: take the expectation on both sides of $N_j = \sum_{i=1}^N W_i(j)$ to find each $\mathbb{E}W_i(j) = N_j/N$. This shows each $W_i(j) \sim \text{Bernoulli}(N_j/N)$, so $\text{Var}(W_i(j)) = N_j(N - N_j)/N^2$. Next, note

$$0 = \text{Cov}(W_1(j), N_j) = \text{Cov}\left(W_1(j), \sum_{i=1}^N W_i(j)\right) = \text{Var}(W_1(j)) + (N-1)\,\text{Cov}(W_1(j), W_2(j))$$

so $\text{Cov}(W_1(j), W_2(j)) = -\text{Var}(W_1(j))/(N-1)$. For $j \neq k$, we have $\text{Cov}(W_1(j), W_1(k)) = \mathbb{E}W_1(j)W_1(k) - (N_j/N)(N_k/N) = -N_jN_k/N^2$ since at least one of $W_i(j), W_i(k)$ is zero. Finally,

$$\begin{aligned} 0 =& \text{Cov}(W_1(j), N_k) = \text{Cov}\left(W_1(j), \sum_{i=1}^N W_i(k)\right) \\ =& \text{Cov}(W_1(j), W_1(k)) + (N-1)\,\text{Cov}(W_1(j), W_2(k)) \end{aligned}$$

so $\text{Cov}(W_1(j), W_2(k)) = -\text{Cov}(W_1(j), W_1(k))/(N-1)$.

Part 2: Aided by (B.1) and part 1, we have

$$\begin{aligned} \text{Cov}(\hat{\bar{x}}(j), \hat{\bar{y}}(j)) =& \frac{1}{N_j^2}\,\text{Cov}\left(\sum_{i=1}^N W_i(j)x_i, \sum_{i=1}^N W_i(j)y_i\right) = \frac{1}{N_j^2} x^\top \,\text{Cov}\begin{pmatrix} W_1(j) \\ \vdots \\ W_N(j) \end{pmatrix} y \\ =& \frac{N_j(N-N_j)}{N(N-1)} x^\top V_N y = \frac{N_j(N-N_j)}{N} S_{xy} \end{aligned}$$

Using $y \leftarrow x$ in the above gives us $\text{Var}(\hat{\bar{x}}(j))$. Similarly, when $j \neq k$

$$\begin{aligned} \text{Cov}(\hat{\bar{x}}(j), \hat{\bar{y}}(k)) =& \frac{1}{N_jN_k}\,\text{Cov}\left(\sum_{i=1}^N W_i(j)x_i, \sum_{i=1}^N W_i(k)y_i\right) \\ =& \frac{1}{N_jN_k} x^\top \,\text{Cov}\left(\begin{pmatrix} W_1(j) \\ \vdots \\ W_N(j) \end{pmatrix}, \begin{pmatrix} W_1(k) \\ \vdots \\ W_N(k) \end{pmatrix}\right) y \\ =& \frac{1}{N_j} N_k \frac{-N_jN_k}{N(N-1)} x^\top V_N y = -S_{xy}/N \end{aligned}$$

Using $y \leftarrow x$ in the above gives us $\text{Cov}(\hat{\bar{x}}(j), \hat{\bar{x}}(k))$.

Part 3: Note

$$\begin{aligned} (N_j - 1)\hat{S}_{xy}(j) =& \sum_{i=1}^N W_i(j)\{x_i - \hat{\bar{x}}(j)\}\{y_i - \hat{\bar{y}}(j)\} \\ =& \sum_{i=1}^N W_i(j)(x_i - \bar{x})(y_i - \bar{y}) - N_j\{\hat{\bar{x}}(j) - \bar{x}\}\{\hat{\bar{y}}(j) - \bar{y}\} \end{aligned}$$

and taking the expectation of both sides

$$\begin{aligned}(N_j-1)\mathbb{E}\hat{S}_{xy}(j) =& \mathbb{E}W_1(j)\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) - N_j\mathbb{E}\{\hat{\bar{x}}(j)-\bar{x}\}\{\hat{\bar{y}}(j)-\bar{y}\}\\ =& \frac{N_j}{N}(N-1)S_{xy} - N_j\operatorname{Cov}(\hat{\bar{x}}(j),\hat{\bar{y}}(j))\\ =& \Big\{\frac{N_j}{N}(N-1) - N_j\frac{N-N_j}{N\cdot N_j}\Big\}S_{xy} = (N_j-1)S_{xy}\end{aligned}$$

where the last step uses part 2. Using $y \leftarrow x$ in the above shows $\mathbb{E}\hat{S}_x^2(j) = S_x^2$. □

In order to prove Lemma 2, we use the next definition and result, which come from [11]. The proof of Lemma 2 is found in [10]. For completeness we have provided their proofs while being sloppier with constants.

Definition 1. If X is a random variable with $\mathbb{E}X = 0$, and $v > 0$, then X is *v-sub-Gaussian* if $\mathbb{E}e^{uX} \le \exp(u^2v/2)$, for all $u \in \mathbb{R}$.

It is thus immediate that, if $0 < v_1 \le v_2$ and X is v_1-sub-Gaussian, then it is also v_2-sub-Gaussian. We collect various other properties of sub-Gaussian random variables:

Lemma 10.
- *If $X_1, \ldots, X_N$ are independent and each X_i is v_i-sub-Gaussian, then $\sum_{i=1}^N X_i$ is $\sum_{i=1}^N v_i$-sub-Gaussian.*
- (sub-Gaussian tail bound) *If X is v-sub-Gaussian, then $\mathbb{P}(X > t) \vee \mathbb{P}(X < -t) \le \exp(-t^2/2v)$*
- *If X is a bounded random variable with $\mathbb{E}X = 0$ and $X \in [a,b]$, then X is $(b-a)^2/4$-sub-Gaussian.*

Proof of Lemma 2. All sub-Gaussian statements in this proof follow from Lemma 10. Without loss of generality, $\bar{Y} = 0$. We first assume $N_1 \le N/2$. By the division algorithm, we can pick $m_1 \in \mathbb{Z}, r \in \{0, 1, \ldots, N_1 - 1\}$ such that $N = N_1m_1 + r$. Note $m_1 = \lfloor N/N_1 \rfloor \ge 2$. We first claim there exists $C_1 \subseteq \{1, \ldots, N\}$ random but with fixed size $\operatorname{card}(C_1) \le N/2$ and $d_1 \in \{\pm 1\}$ such that

$$\mathbb{E}\exp\big(u\sum_{i\in A}Y_i\big) \le \mathbb{E}\exp\big(ud_1\sum_{i\in C_1}\frac{Y_i}{m_1(m_1+1)} + \frac{NSu^2}{4}\big) \tag{B.2}$$

for all $u \in \mathbb{R}$, where d_1 is not random because it depends only on N,N_1.

To show this, we consider the following scheme to generate a SRS of size N_1 from $\{1, \ldots, N\}$. Pick $\pi \sim \operatorname{Unif}(\Pi_N)$, partition it into consecutive blocks $B_1, \ldots, B_{N_1}$, where $\operatorname{card}(B_k) = m_1 + 1$ for $k = 1, \ldots, r$ and $\operatorname{card}(B_k) = m_1$ for $k = r+1, \ldots, N_1$. Then for $k = 1, \ldots, N_1$, pick $w_k \sim \operatorname{Unif}(B_k)$ independently across different k, and put $A := \{w_1, \ldots, w_{N_1}\}$. For convenience, let $\bar{Y}_k$ denote the mean of $\{Y_i : i \in B_k\}$, $B := \cup_{k=1}^r B_k$.

We first condition on π. Then $Y_{w_k} - \bar{Y}_k$ is contained in an interval whose squared length is upper bounded by

$$\max_{i,j\in B_k}(Y_i - Y_j)^2 \le 2\max_{i,j\in B_k}(Y_i^2 + Y_j^2) \le 2\sum_{i\in B_k} Y_i^2$$

so $Y_{w_k} - \bar{Y}_k$ is $\sum_{i\in B_k} Y_i^2/2$-sub-Gaussian, and $\sum_{k=1}^{N_1}(Y_{w_k} - \bar{Y}_k)$ is $\sum_{i=1}^{N_1} Y_i^2/2$-sub-Gaussian, hence $NS/2$-sub-Gaussian, as $\sum_{i=1}^{N_1} Y_i^2 = (N-1)S \le NS$. Then

$$\begin{aligned}\mathbb{E}\exp\big(u\sum_{i\in A} Y_i\big|\pi\big) &= \exp\big(u\sum_{k=1}^{N_1}\bar{Y}_k\big)\mathbb{E}\big\{\exp\big(u\sum_{k=1}^{N_1}(Y_{w_k} - \bar{Y}_k)\big)\big|\pi\big\} \\ &\le \exp\big(u\sum_{k=1}^{N_1}\bar{Y}_k + \frac{NSu^2}{4}\big) = \exp\big(\frac{-u}{m_1(m_1+1)}\sum_{i\in B} Y_i + \frac{NSu^2}{4}\big)\end{aligned}$$

where for the $\le$ we use Definition 1, and for the second equality we note that $\sum_{k=1}^{N_1}\bar{Y}_k = (m_1+1)^{-1}\sum_{k=1}^{r}\sum_{i\in B_k} Y_i + m_1^{-1}\sum_{k=r+1}^{N_1}\sum_{i\in B_k} Y_i = -\{m_1(m_1+1)\}^{-1}\sum_{i\in B} Y_i$, using $\sum_{i=1}^{N} Y_i = 0$. Thus, by the tower property for conditional expectation

$$\mathbb{E}\exp\big(u\sum_{i\in A} Y_i\big) \le \mathbb{E}\exp\big(\frac{-u}{m_1(m_1+1)}\sum_{i\in B} Y_i + \frac{NSu^2}{4}\big)$$

Now if $\mathrm{card}(B) \le N/2$, then we take $C_1 = B$ and $d_1 = -1$. Else, noting $\sum_{i\in B^c} Y_i = -\sum_{i\in B} Y_i$, we take $C_1 = B^c$ and $d_1 = +1$, which shows (B.2). Note d_1 depends only on $\mathrm{card}(B)$, which in turn only depends on N, N_1.

Applying (B.2) to itself, and considering the population $\{Y_i/m_1(m_1+1) : i = 1,\dots,N\}$ which has mean zero and variance $S/\{m_1(m_1+1)^2\} \le S/36$, we have that there exists $C_2 \subseteq \{1,\dots,N\}$ random but with $\mathrm{card}(C_2) \le N_2$ and $d_2 \in \{\pm 1\}$ depending on $N, N_1, \mathrm{card}(C_1)$ such that

$$\mathbb{E}\exp\big(u\sum_{i\in A} Y_i\big) \le \mathbb{E}\exp\big\{\frac{ud_2}{m_2(m_2+1)m_1(m_1+1)}\sum_{i\in C_2} Y_i + (1+\frac{1}{36})\frac{NSu^2}{4}\big\}$$

By an inductive argument, to each $k \in \mathbb{Z}^+$ corresponds $C_k \in \{1,\dots,N\}$ random but with $\mathrm{card}(C_k) \le N/2$ fixed, and $d_k \in \{\pm 1\}$ depending on $N, N_1, \mathrm{card}(C_1), \dots, \mathrm{card}(C_{k-1})$ such that

$$\begin{aligned}\mathbb{E}\exp\big(u\sum_{i\in A} Y_i\big) &\le \mathbb{E}\exp\big\{\frac{ud_k}{\prod_{j=1}^{k} m_j(m_j+1)}\sum_{i\in C_k} Y_i + \sum_{j=0}^{k-1}\frac{1}{36^j}\frac{NSu^2}{4}\big\} \\ &\le \exp\big\{\frac{|u|}{\prod_{j=1}^{k} m_j(m_j+1)}\sum_{i=1}^{N}|Y_i| + \frac{36}{35}\frac{NSu^2}{4}\big\} \\ &\le \exp\big(\frac{N|u|\sqrt{S}}{6^k} + \frac{36}{35}\frac{NSu^2}{4}\big)\end{aligned}$$

$$\text{so } \mathbb{E}\exp\big(u\sum_{i\in A} Y_i\big) \le \exp\big(\frac{36}{35}\frac{NSu^2}{4}\big) \tag{B.3}$$

where the last line is because $k \in \mathbb{Z}^+$ is arbitrary. If $\text{card}(A) > N/2$, then (B.3) holds with A^c in place of A, and we simply use $\sum_{i\in A^c} Y_i = -\sum_{i\in A} Y_i$ to replace A^c back with A.

Also from (B.3), $\sum_{i\in A} Y_i$ is $18NS/35$-sub-Gaussian, from which we conclude, for $t \ge 0$

$$\mathbb{P}(\hat{\hat{Y}} \ge t) \vee \mathbb{P}(\hat{\hat{Y}} \le -t) = \mathbb{P}\big(\sum_{i\in A} Y_i \ge N_1 t\big) \vee \mathbb{P}\big(\sum_{i\in A} Y_i \le -N_1 t\big) \le \exp\big(\frac{-35N_1^2}{36NS}t^2\big) \qquad \square$$

We next provide an additional definition and lemma relevant to proving Proposition 4.

Definition 2 (inverse)**.** If $J \subseteq \mathbb{R}$ is an open interval (possibly unbounded), $f : J \to \mathbb{R}$ is nondecreasing, then we define $f^{-1}(u) := \inf\{x \in J : f(x) \ge u\}$ for $u \in \mathbb{R}$. If J has finite left and/or right endpoints a, b, respectively, regard for convenience $f(a) = -\infty$, $f(b) = \infty$, so that $f^{-1}(u) \in J$ for $u \in \mathbb{R}$.

An immediate consequence of the above definition that we use repeatedly is $u \le f(x)$ if and only if $f^{-1}(u) \le x$. Equivalently, by negating, $u > f(x)$ if and only if $f^{-1}(u) > x$.

Lemma 11. *Let X, Y be random variables where $X \le_{\text{st}} Y$, put $F_X(x) = \mathbb{P}(X \le x)$, $G_X(x) = \mathbb{P}(X < x)$, and define F_Y, G_Y similarly. Let $U \sim \text{Unif}(0,1)$.*

- *$F_X^{-1}(U) \overset{\text{d}}{=} G_X^{-1}(U) \overset{\text{d}}{=} X$.*
- *$F_X(x) = \lim_{t\downarrow x} G_X(t)$ and $G_X(x) = \lim_{t\uparrow x} F_X(t)$. In other words, F_X is determined by G_X and vice versa. As a consequence, $X \le_{\text{st}} Y$ if and only if $G_Y \le G_X$ on $\mathbb{R}$.*
- *If f is a real valued, nondecreasing function, then $f(X) \le_{\text{st}} f(Y)$.*
- *$G_X(X) \le_{\text{st}} U \le_{\text{st}} F_X(X)$.*

Proof. Part 1 follows readily by noting $\mathbb{P}\{F_X^{-1}(U) \le x\} = \mathbb{P}\{U \le F_X(x)\} = F_X(x)$, hence $F_X^{-1}(U)$ has the same CDF as X. Similarly, $\mathbb{P}\{G_X^{-1}(U) > x\} = \mathbb{P}\{U \ge G_X(t)\} = 1 - G_X(t)$. Subtracting both sides from 1 shows $G_X^{-1}(U)$ has the same CDF as X.

Part 2 is easy. For part 3, note $\mathbb{P}\{f(X) \ge x) = \mathbb{P}\{A \ge f^{-1}(x)\} \le \mathbb{P}\{B \ge f^{-1}(x)\} = \mathbb{P}\{f(B) \ge x\}$. For part 4, the desired result is equivalent to (by definition and part 1)

$$\begin{aligned} \mathbb{P}\{F_X(X) \le x\} \le & x \le \mathbb{P}\{G_X(X) \le x\} \iff \\ \mathbb{P}\{F_X \circ F_X^{-1}(U) \le x\} \le & \mathbb{P}(U \le x) \le \mathbb{P}\{G_X \circ G_X^{-1}(U) \le x\} \end{aligned}$$

which follows because for $u \in (0,1)$, $G_X \circ G_X^{-1}(u) \le u \le F_X \circ F_X^{-1}(u)$. Indeed, for $x > F_X^{-1}(u)$ and $t < G_X^{-1}(u)$, we have $G_X(t) < u \le F_X(x)$. Now take $t \uparrow G_X^{-1}(u)$ and $x \downarrow F_X^{-1}(u)$, and use the left (resp right) continuity of G_X (resp F_X), we have $u \in (0,1)$, $G_X \circ G_X^{-1}(u) \le u \le F_X \circ F_X^{-1}(u)$. $\square$

B.2 Computational Details of the Simulations

ANOVA and Factorial. As we remarked in the main text, under balanced designs and in an ANOVA or factorial designs setting, $B = F$ by Proposition 5. Indeed, in the ANOVA setting, we have $C = (1_J \ -I_J)$. Note $\mathcal{C}(C^\top)^\perp = \mathcal{C}(1_J)$, so the projection matrix onto $\mathcal{C}(C^\top)$ must be $M = I_J - 1_J 1_J^\top / J$. In factorial designs, C has orthogonal rows and each entry of C is ± 1, hence $M \propto C^\top C$. Thus, in both ANOVA and factorial designs, M has equal entries on its main diagonal. Going forward, we consider B only.

To investigate situations where B may fail, we invoke Theorem 3 rather than Theorem 2, since it involves eigenvalues of a lower dimensional matrix. Under $H_{0\mathrm{N}}(C, x)$, we have

$$m \cdot B \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} \lambda_j \big(CVC^\top (\bar{S} C P^{-1} C^\top)^{-1}\big) \xi_j^2, \ m \cdot B_\pi | W \xrightarrow{\mathrm{d}} \chi^2_m$$

We thus have cause to consider the eigenvalues of the matrix

$$CVC^\top (\bar{S} C P^{-1} C^\top)^{-1} = CVC^\top (CC^\top)^{-1} / \sum_{j=1}^{J} S^2(j)$$

because in a balanced design $\bar{S} = \sum_{j=1}^{J} S(j,j)/J$ and $P = I_J / J$ from.

For further exploration, to avoid tedious algebra, we consider the special case that the potential outcomes have rank one covariance structure $S = uu^\top$, or approximately so. This still leaves J degrees of freedom in picking the entries of u. We also considered a diagonal covariance structure for S, but this turned out to be more restrictive, despite still having J degrees of freedom to pick the main diagonal of S. When $S = uu^\top$, each $S(j,j) = u_j^2$. Recalling what V is from Proposition 3, we have

$$CVC^\top (\bar{S} C P^{-1} C^\top)^{-1} = \frac{1}{|u|^2} C \big(J \cdot \mathrm{diag}(u_1^2, \ldots, u_J^2) - uu^\top\big) C^\top (CC^\top)^{-1}$$

At this point we begin to consider ANOVA and factorial designs separately. We take $J = 3$ for ANOVA and $K = 2$ (hence $J = 4$) for factorial designs, where in the latter we test for no main effects. This gives us the contrast matrices

$$C_A = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \ C_F = \begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix}$$

and carrying out the matrix algebra gives

$$\begin{aligned} &C_A V C_A^\top (\bar{S} C_A P^{-1} C_A^\top)^{-1} \\ &\quad = \frac{1}{3|u|^2} \begin{pmatrix} 3(u_1^2 + u_2^2) - (u_1 - u_2)^2 & 3u_1^2 - (u_1 - u_2)(u_1 - u_3) \\ 3u_1^2 - (u_1 - u_2)(u_1 - u_3) & 3(u_1^2 + u_3^2) - (u_1 - u_3)^2 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \\ &C_F V C_F^\top (\bar{S} C_F P^{-1} C_F^\top)^{-1} = \frac{1}{4|u|^2} \begin{pmatrix} 4|u|^2 - (g_1^\top u)^2 & 4g_3^\top u - (g_1^\top u)(g_2^\top u) \\ 4g_3^\top u - (g_1^\top u)(g_2^\top u) & 4|u|^2 - (g_2^\top u)^2 \end{pmatrix} \end{aligned}$$

where we recall that $g_1^\top$=(-1 -1 1 1), $g_2^\top$=(-1 1 1 -1), and $g_3^\top$=(1 -1 -1 1). For ANOVA, take $u^\top$=(1 2 3) and for factorial designs take $u^\top$=(3 1 1 3). Then the matrices above are

$$\frac{1}{42}\begin{pmatrix}14 & 1\\ 1 & 26\end{pmatrix},\ \frac{1}{5}\begin{pmatrix}5 & 4\\ 4 & 5\end{pmatrix}$$

Using their eigenvalues, we thus have that the permutation distributions of X^2 and $2B$ are both asymptotically χ^2_2 (regardless of whether ANOVA or factorial designs) and under $H_{0\mathrm{N}}(C,x)$ we get (6.1):

$$\begin{aligned} X^2 &\xrightarrow{\mathrm{d}} \xi_1^2 + 0.758\xi_2^2, \qquad 2B \xrightarrow{\mathrm{d}} 1.423\xi_1^2 + 0.434\xi_2^2, && \text{(ANOVA)},\\ X^2 &\xrightarrow{\mathrm{d}} \xi_1^2 + \xi_2^2 \overset{\mathrm{d}}{=} \chi^2_2, \qquad 2B \xrightarrow{\mathrm{d}} 1.8\xi_1^2 + 0.2\xi_2^2, && \text{(Factorial)}. \end{aligned}$$

The calculation of the weights of the asymptotic distribution of X^2 is omitted, but is easily done using Theorem 1. As guaranteed in the main text, each weight for X^2 is ≤ 1, while the weights for $2B$ are merely ≤ 1 on average. In fact, in our factorial designs example, the FRT with X^2 is not even conservative, since the asymptotic distribution of X^2 is exactly χ^2_2, not stochastically dominated by it.

In our examples, the FRT with B will asymptotically fail to control type I error. One way to see this is to compute the variance. The permutation distribution has asymptotic variance $\mathrm{Var}(\chi^2_2) = 4$, but the ranodmization distributions under Neyman's null have asymptotic variances $2(1.423^2 + 0.434^2)$=4.427 and $2(1.8^2 + 0.2^2)$=6.56, in the ANOVA and factorial designs setting, respectively. It is also remarkable that settings with $J > 3$ are not needed to illustrate the failure of B.

To speed up computation, we should exploit the special structure of the designs, as the general forms of the statistics involve matrix multiplications. This is especially helpful as the FRT involves repeated recalculation of these statistics. For ANOVA, we would recommend using (4.2) to compute X^2. Since the projection matrix onto the row space of C is $M = I_J - 1_J 1_J^\top/J$, the Box-type statistic (3.2) simplifies to

$$B = \frac{J\sum_{j=1}^J \hat{\bar{Y}}(j)^2 - \{\sum_{j=1}^J \hat{\bar{Y}}(j)\}^2}{(J-1)\sum_{j=1}^J \hat{S}(j,j)/N_j} = \frac{N[\sum_{j=1}^J \hat{\bar{Y}}(j)^2 - \{\sum_{j=1}^J \hat{\bar{Y}}(j)\}^2/J]}{(J-1)\sum_{j=1}^J \hat{S}(j,j)}$$

where the second equality is due to the balanced design. For factorial designs, in the particular case of $C = C_F$, we have

$$\begin{aligned} X^2 =& \frac{(\hat\tau_1 - \hat\tau_2)^2}{\hat S(2,2)/N_2 + \hat S(3,3)/N_3} + \frac{(\hat\tau_1 + \hat\tau_2)^2}{\hat S(1,1)/N_1 + \hat S(4,4)/N_4}\\ =& \frac{\{\hat{\bar{Y}}(3) - \hat{\bar{Y}}(2)\}^2}{\hat S(2,2)/N_2 + \hat S(3,3)/N_3} + \frac{\{\hat{\bar{Y}}(4) - \hat{\bar{Y}}(1)\}^2}{\hat S(1,1)/N_1 + \hat S(4,4)/N_4}\\ =& \frac{N}{4}\Big[\frac{(\hat\tau_1 - \hat\tau_2)^2}{\hat S(2,2) + \hat S(3,3)} + \frac{(\hat\tau_1 + \hat\tau_2)^2}{\hat S(1,1) + \hat S(4,4)}\Big]\\ B =& \frac{\{\hat{\bar{Y}}(3) - \hat{\bar{Y}}(2)\}^2 + \{\hat{\bar{Y}}(4) - \hat{\bar{Y}}(1)\}^2}{\sum_{j=1}^4 \hat S(j,j)/N_j} = \frac{N}{4}\frac{\{\hat{\bar{Y}}(3) - \hat{\bar{Y}}(2)\}^2 + \{\hat{\bar{Y}}(4) - \hat{\bar{Y}}(1)\}^2}{\sum_{j=1}^4 \hat S(j,j)} \end{aligned}$$

where the last equalities for X^2 and B are due to the balanced design.

As a final comment, our simulations build on the ones found in [24]. For one thing, we have also considered factorial designs. For another, we have demonstrated an ANOVA example where the FRT with B may fail even with balanced designs. In [24], both diagonal and rank one covariance structures were considered for the potential outcomes, though only the former in balanced designs. We illustrated an interesting situation when the latter is considered, even with the same marginal variances.

ANOVA and SRE. We discuss some computational issues that arise with SRE's. From (3.1) and (5.1), the X^2 for a SRE differs from that for a CRE by $\breve{Y}$ and $\sum_{h=1}^{H} N_{[h]}\hat{D}_{[h]}/N$ in place of $\hat{\bar{Y}}$ and $\hat{D}$, respectively. Because of the ANOVA setup, we can still depend on (4.2) for a faster computation of X^2. All we have to do is replace all $\hat{\bar{Y}}(j)$ by $\sum_{h=1}^{H} N_{[h]}\hat{\bar{Y}}_{[h]}(j)$ and all $\hat{S}(j,j)/N_j$ by

$$\sum_{h=1}^{H} \frac{N_{[h]}^2}{N^2}\frac{\hat{S}(j,j)}{N_{[h]j}}$$

In the simulations of the main text, we size up X^2 from SRE against the more common F statistic from linear regression of the observed response on stratum and treatment indicators. This is nominally $J+H$ predictors, but $J+H-1$ linearly independent ones. The computation of this F statistic becomes more tractable due to the balanced design $N_{[h]j} = N/(HJ)$. The well-known formula (eg from [18]) in this case is

$$\begin{aligned} F =& \frac{HN_{[1]1}\sum_{j=1}^{J}(\bar{Y}_{\bullet j}^{\text{obs}} - \bar{Y}_{\bullet\bullet}^{\text{obs}})^2/(J-1)}{\sum_{i,j,h} 1(W_i = j, X_i = h)(Y_i^{\text{obs}} - \bar{Y}_{\bullet j}^{\text{obs}} - \bar{Y}_{[h]\bullet}^{\text{obs}} + \bar{Y}_{\bullet\bullet}^{\text{obs}})^2/(N-H-J+1)} \\ =& \frac{\text{SS}_{\text{Tmt}}/(J-1)}{\text{SS}_{\text{E}}/(N-H-J+1)} \end{aligned}$$

where $\bar{Y}_{\bullet j}^{\text{obs}} = \sum_{i=1}^{N} 1(W_i = j)Y_i^{\text{obs}}/HN_{[1]1}$ are the means for fixed treatments, $\bar{Y}_{[h]\bullet}^{\text{obs}} = \sum_{i=1}^{N} 1(X_i = h)Y_i^{\text{obs}}/JN_{[1]1}$ are the means for fixed strata, and $\bar{Y}_{\bullet\bullet}^{\text{obs}} = \sum_{i=1}^{N} Y_i^{\text{obs}}/N$ is the grand mean. Our notation is intentionally reminiscent of that from randomized complete block designs.

Another computational shortcut is the sum of squares decomposition. If we define the within-stratum sum of squares $\text{SS}_{\text{Bl}} = JN_{[1]1}\sum_{h=1}^{H}(\bar{Y}_{[h]\bullet}^{\text{obs}} - \bar{Y}_{\bullet\bullet}^{\text{obs}})^2$, then $\text{SS}_{\text{E}} = \sum_{i=1}^{N}(Y_i^{\text{obs}} - \bar{Y}_{\bullet\bullet}^{\text{obs}})^2 - \text{SS}_{\text{Tmt}} - \text{SS}_{\text{Bl}}$. The identity $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$ also saves some computational effort. A final note is that our computational shortcut for X^2 extends to imbalanced SRE's, but that for F does not. Adding any constant effect between the two strata produces identical results because all test statistics we consider subtract the sample mean at some point.

Theorem 4 tells us that X^2 has the same asymptotic distribution as listed in (6.1). Obtaining the asymptotic distribution of F under $H_{0\text{N}}$ in a blocked experiment is beyond our present scope. However, given our particular setup, intuition suggests that the blocked version of F also has the same asymptotic distribution given in (6.1). Another takeaway is,

despite the pains taken to ensure the settings of the ANOVA simulation were as similar as possible to its blocked counterpart, their behavior was still somewhat different.

In our simulation, N =15,60,120 for CRE ANOVA, N =20,80,160 for Factorial, and N=30,120,240 for SRE ANOVA. The sample size of SRE ANOVA is effectively twice that of CRE ANOVA, so we expect better asymptotics in the former. Another way to see this is that, roughly speaking, the "SRE" X^2 in (5.1) averages the results of two independent "CRE" X^2 values because each stratum in SRE ANOVA has identical potential outcomes, except for a location shift, with CRE ANOVA.

Because a balanced design and normally generated potential outcomes is in some sense the most favorable situation imaginable, we do not advocate the FRT with X^2 as a test for the weak null when the sample size is very small because of the mediocre simulation results. When $J = 3$, the balanced design still offers much (but not sufficient) protection against heteroscedasticity. When $J = 4$, we have more flexibility for more adverse parameters.

Confidence Regions. We consider a balanced factorial design with $K = 2$, and each $N_j = 10$. Potential outcomes are generated so that $\bar{Y}(1) = \ldots = \bar{Y}(J)$. In particular, $\tau_1 = \tau_2 = 0$, and there is no average interaction effect. With all groups having the same mean, we can assess type I error. For the simulation, we also "know" $\tilde{x} = 0_{J-m-1}$, avoiding that sticky issue. The noise has distribution $U^2 - 1/3$, where $U \sim \text{Unif}(0, 1)$. We then multiply each Y_i by the same 4×4 matrix to give the groups different variances:

$$\begin{pmatrix} 2 & 1 & 3/2 & 1 \\ 0 & \sqrt{5} & \sqrt{5}/2 & 2/\sqrt{5} \\ 0 & 0 & 3/\sqrt{2} & 1/\sqrt{2} \\ 0 & 0 & 0 & \sqrt{3.7} \end{pmatrix}, \text{ so } S \propto \begin{pmatrix} 4 & 2 & 3 & 2 \\ 2 & 6 & 4 & 3 \\ 3 & 4 & 8 & 4 \\ 2 & 3 & 4 & 6 \end{pmatrix}$$

The former matrix is in fact the Cholesky decomposition of the latter matrix. We purposely designed the potential outcomes to be skewed and used a small sample size, since otherwise FRT and asymptotic results were indistinguishable.

As stated, we have three situations for which we want confidence regions: τ_1 and τ_2 individually and jointly. This corresponds to contrast matrices

$$C_1 = \text{(-1 -1 1 1)}, \; C_2 = \text{(-1 1 -1 1)}, \; C_{12} = \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}$$

The X^2 statistic (3.1) becomes in these cases

$$X^2 = N\frac{(\hat{\tau}_1 - x)^2}{\sum_{j=1}^4 \hat{S}(j)} \text{ to test } \tau_1 = x, \; X^2 = N\frac{(\hat{\tau}_2 - x)^2}{\sum_{j=1}^4 \hat{S}(j,j)} \text{ to test } \tau_2 = x$$

$$X^2 = \frac{N}{4}\Big(\frac{(\hat{\tau}_1 - x_1 + \hat{\tau}_2 - x_2)^2}{\hat{S}(1,1) + \hat{S}(4,4)} + \frac{(\hat{\tau}_1 - x_1 - \hat{\tau}_2 + x_2)^2}{\hat{S}(2,2) + \hat{S}(3,3)}\Big) \text{ to test } \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

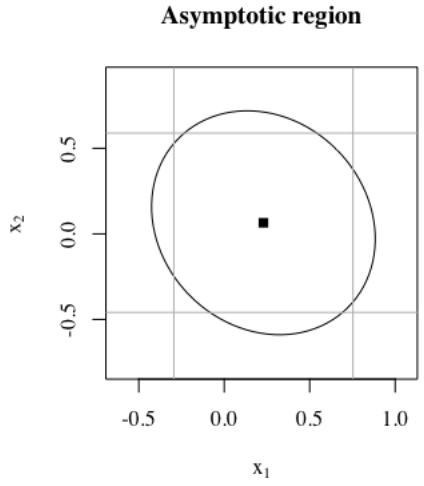

Figure B.1: For our simulated data, the asymptotic 0.95 joint confidence region for τ_1, τ_2. The vertical and horizontal lines are the endpoints of the 0.95 CI for τ_1, τ_2, respectively. The point ($\hat{\tau}_1$,$\hat{\tau}_2$) is also shown. The entire plotting region corresponds to our search region.

We must search over a grid of possible (x_1, x_2) values. To narrow down the search region, we first compute the asymptotic confidence regions (4.3):

$$\text{For } \tau_1 : \{x : X^2 \leq \chi^2_{1,\alpha}\} = \big[\hat{\tau}_1 - \big(\frac{\sum_{j=1}^4 \hat{S}(j,j)}{N}\chi^2_{1,\alpha}\big)^{1/2}, \hat{\tau}_1 + \big(\frac{\sum_{j=1}^4 \hat{S}(j,j)}{N}\chi^2_{1,\alpha}\big)^{1/2}\big]$$

$$\text{For } \tau_2 : \{x : X^2 \leq \chi^2_{1,\alpha}\} = \big[\hat{\tau}_2 - \big(\frac{\sum_{j=1}^4 \hat{S}(j,j)}{N}\chi^2_{1,\alpha}\big)^{1/2}, \hat{\tau}_2 + \big(\frac{\sum_{j=1}^4 \hat{S}(j,j)}{N}\chi^2_{1,\alpha}\big)^{1/2}\big]$$

Coincidentally, the asymptotic width of the CI's for the main effects is always the same, and this still holds for factorial designs with $K > 2$, and no matter what the estimated group variances $\hat{S}(j,j)$ are. The joint confidence region is asymptotically an ellipse centered at $(\hat{\tau}_1, \hat{\tau}_2)$, with one axis in the $(1,1)$ direction with radius $2[\{\hat{S}(1,1) + \hat{S}(4,4)\}\chi^2_{2,\alpha}/N]^{1/2}$ and the other axis in the $(1,-1)$ direction with radius $2[\{\hat{S}(2,2) + \hat{S}(3,3)\}\chi^2_{2,\alpha}/N]^{1/2}$. For concreteness, we pick test level $\alpha = 0.05$. We search over a rectangular (in fact square) region that will be slightly wider than the ellipse as an added precaution.

See Figure B.1 for the realization of the asymptotic regions for our simulated data. Note that, at level α=0.05, there exists x_1, x_2 such that either $\tau_1 = x_1$ or $\tau_2 = x_2$ are rejected, while $(\tau_1, \tau_2) = (x_1, x_2)$ is not rejected. But there does not exist x_1, x_2 where both marginals are rejected while the joint hypothesis is not rejected. The grid resolution will be 55 points for both x_1, x_2, so 55^2=3025 hypotheses $\tau_1 = x_1$, $\tau_2 = x_2$ need to be tested by the FRT. To test for nonzero x, we must impute potential outcomes, and in our three cases FRT-2

simplifies to

$$\text{For } \tau_1 = x : Y_i^* = \begin{pmatrix} Y_i^{\text{obs}} \\ Y_i^{\text{obs}} \\ Y_i^{\text{obs}} + x \\ Y_i^{\text{obs}} + x \end{pmatrix} 1(W_i \in \{1,2\}) + \begin{pmatrix} Y_i^{\text{obs}} - x \\ Y_i^{\text{obs}} - x \\ Y_i^{\text{obs}} \\ Y_i^{\text{obs}} \end{pmatrix} 1(W_i \in \{3,4\})$$

$$\text{For } \tau_2 = x : Y_i^* = \begin{pmatrix} Y_i^{\text{obs}} \\ Y_i^{\text{obs}} + x \\ Y_i^{\text{obs}} \\ Y_i^{\text{obs}} + x \end{pmatrix} 1(W_i \in \{1,3\}) + \begin{pmatrix} Y_i^{\text{obs}} - x \\ Y_i^{\text{obs}} \\ Y_i^{\text{obs}} - x \\ Y_i^{\text{obs}} \end{pmatrix} 1(W_i \in \{2,4\})$$

$$\text{For } \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : Y_i^* = \begin{pmatrix} Y_i^{\text{obs}} \\ Y_i^{\text{obs}} + x_2 \\ Y_i^{\text{obs}} + x_1 \\ Y_i^{\text{obs}} + x_1 + x_2 \end{pmatrix} 1(W_i = 1) + \begin{pmatrix} Y_i^{\text{obs}} - x_2 \\ Y_i^{\text{obs}} \\ Y_i^{\text{obs}} + x_1 - x_2 \\ Y_i^{\text{obs}} + x_1 \end{pmatrix} 1(W_i = 2)$$
$$+ \begin{pmatrix} Y_i^{\text{obs}} - x_1 \\ Y_i^{\text{obs}} + x_2 - x_1 \\ Y_i^{\text{obs}} \\ Y_i^{\text{obs}} + x_2 \end{pmatrix} 1(W_i = 3) + \begin{pmatrix} Y_i^{\text{obs}} - x_1 - x_2 \\ Y_i^{\text{obs}} - x_1 \\ Y_i^{\text{obs}} - x_2 \\ Y_i^{\text{obs}} \end{pmatrix} 1(W_i = 4)$$

In all cases we used the completed matrix

$$\begin{pmatrix} C \\ \tilde{C} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

which, being symmetric and orthogonal, is its own inverse. Through this process, we ultimately get, at each (x_1, x_2) in the grid, a set of imputed potential outcomes.

For each of 2500 permutations of $\{1, \ldots, N\}$, we compute X_π^2 at each of the 55^2 grid points for the joint test, and at each of the $2 \cdot 55$ grid points for the two marginal tests. It is important, for each pass through the grid, to use the same permutation, for smoothness reasons. In the one-dimensional case, it increases the chance that the confidence set will in fact be an interval. The p-value is then calculated at each grid point.

B.3 More about treatment-control

We are now about to prove Corollary 3. The proof has some interesting results in their own right about the treatment-control setting. We lay these out before the actual proof.

Recall that, in the treatment-control setting, $J = 2$, and unit i either receives the treatment (then $Y_i^{\text{obs}} = Y_i(1)$) or control (then $Y_i^{\text{obs}} = Y_i(2)$). A parameter of interest is the average treatment effect $\tau = \bar{Y}(1) - \bar{Y}(2)$. The weak null hypothesis is $H_{0\text{N}} = H_{0\text{N}}(C, 0) : \tau = 0$ where $C = (1, -1)$ is a row vector. Both X^2 and B are proper, and reduce to (the square

of) a statistic proposed in [72]:

$$X^2 = B = \frac{\{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)\}^2}{\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2} = \frac{\hat{\tau}^2}{\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2} \tag{B.4}$$

where we recall $\hat{\tau} = \hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)$ is the sample difference-in-means. This was first proposed by [72] as an unbiased estimator of τ. In conjunction with its intuitive appeal, this explains why $\hat{\tau}$ is a popular statistic. Under $H_{0\mathrm{N}}$ and Assumption A, it satisfies

$$N^{1/2}\hat{\tau} \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \frac{S(1,1)}{p_1} + \frac{S(2,2)}{p_2} - S_\tau^2\big) \tag{B.5}$$

where S_τ^2 is the variance of $\{Y_i(1) - Y_i(2)\}$, i.e. $S_\tau^2 = (N-1)^{-1}\sum_{i=1}^N \{Y_i(1) - Y_i(2) - \bar{Y}(1) + \bar{Y}(2)\}^2$. The variance in (B.5) is *Neyman's variance formula*. We therefore have under $H_{0\mathrm{N}}$

$$X^2 \xrightarrow{\mathrm{d}} a \cdot \chi_1^2, \text{ where } a = \frac{S(1,1)/p_1 + S(2,2)/p_2 - S_\tau^2}{S(1,1)/p_1 + S(2,2)/p_2}, \; X_\pi^2 | W \xrightarrow{\mathrm{d}} \chi_1^2 \tag{B.6}$$

Note $a \in [0,1]$ measures how conservative the FRT with X^2 is: smaller a means more conservative. Unfortunately, this constant a cannot be computed or even approximated in practice because of the presence of $S(1,2)$ in S_τ^2. Observe that if $S(1,2)$ is large and negative, that is the treatment and control values are highly negatively correlated, then $a \approx 0$. On the flip side, if $S(1,2)$ is large and positive, that is the treatment and control values are highly positively correlated, then $a \approx 1$. Taking this to the extreme, under strict additivity, $S(1,1) = S(2,2) = S(1,2)$, in which case we have $a = 1$.

The denominator of (B.4) is $\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2$. Multiplied by N, it was the proposal by [72] as an estimator of the variance in (B.5), to overcome the fact that S_τ^2 is not identifiable. By Lemma 9 part 3, we have that, in expectation, this estimator exceeds the true variance. This ties intimately into (B.6).

Now we turn to statistics that are not proper. Recall that a common statistic in treatment-control is $|\hat{\tau}|$. Note the statistic has an absolute value to reflect that we have a two-sided test. However, this statistic is not proper. A result about the permutation distribution from [23] that we also later derive shows this:

$$N^{1/2}\hat{\tau}_\pi | W \xrightarrow{\mathrm{d}} \mathcal{N}\big(0, \frac{S(1,1)}{p_2} + \frac{S(2,2)}{p_1} + \tau^2\big) \overset{\mathrm{d}}{=} \mathcal{N}\big(0, \frac{S(1,1)}{p_2} + \frac{S(2,2)}{p_1}\big) \text{ under } H_{0\mathrm{N}}$$

Note that, unlike $X^2 = B$ and F, the permutation distribution of $\hat{\tau}$ is not invariant to the veracity of $H_{0\mathrm{N}}$. Comparing with (B.5), the criterion of Proposition 4 is not met. We do not always have, under $H_{0\mathrm{N}}$, that $|N^{1/2}\hat{\tau}| \leq_{\mathrm{st}} |N^{1/2}\hat{\tau}_\pi|$, even asymptotically. Since both are asymtotically normal, we equivalently say the permutation variance V_{F} is not always an upper bound for the randomization variance V_{N}. Indeed, $|\mathcal{N}(0, V_{\mathrm{F}})| \leq_{\mathrm{st}} |\mathcal{N}(0, V_{\mathrm{N}})|$ when $V_{\mathrm{F}} \leq V_{\mathrm{N}}$.

The asymptotic behavior of the F statistic stated in Theorem 3 is also simplified in the treatment-control setting. Because of the equivalence of the F's established in Proposition

6, all results about F found in [24] apply. In particular Corollary 5 found therein states, under $H_{0\mathrm{N}}$

$$F \xrightarrow{\mathrm{d}} C_1\chi_1^2, \text{ where } C_1 := \lim_{N\to\infty} \frac{\mathrm{Var}(\hat{\tau})}{S(1,1)/N_2 + S(2,2)/N_1} = \frac{S(1,1)/p_1 + S(2,2)/p_2 - S_\tau^2}{S(1,1)/p_2 + S(2,2)/p_1}$$

where the limit can be evaluated by (B.5). We also have $F_\pi | W \xrightarrow{\mathrm{d}} \chi_1^2$. Unlike a in (B.6), which is in $[0,1]$, $C_1 > 1$ is possible by the same choice of S^2, p_j from earlier. From this, we see that the F statistic derived from a linear models framework may fail even in the case of treatment-control. In fact, it seems the F statistic fails precisely when $|\hat{\tau}|$ does and vice versa.

Proof of Corollary 3. Because $C = (1,\ -1)$ is a row vector, Corollary 1 immediately gives that $B = X^2$, which is proper. The closed form B.4 is convenient, which we now work out:

$$\begin{aligned}
X^2 =& \{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)\}\big\{(1,\ -1)\,\mathrm{diag}\left(\hat{S}(1,1)/N_1, \hat{S}(2,2)/N_2\right)\begin{pmatrix}1\\-1\end{pmatrix}\big\}^{-1}\{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)\} \\
=& \frac{\{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)\}^2}{\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2} \\
B =& \big\{\frac{1}{2}(\hat{\bar{Y}}(1),\ \hat{\bar{Y}}(2))\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\begin{pmatrix}\hat{\bar{Y}}(1)\\ \hat{\bar{Y}}(2)\end{pmatrix}\big\} \\
& \div \big\{\frac{1}{2}\,\mathrm{tr}(\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}\mathrm{diag}\left(\hat{S}(1,1)/N_1, \hat{S}(2,2)/N_2\right)\big)\big\} \\
=& \frac{\{\hat{\bar{Y}}(1) - \hat{\bar{Y}}(2)\}^2}{\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2}
\end{aligned}$$

where we have used

$$M = C^\top(CC^\top)^{-1}C = \frac{1}{2}\begin{pmatrix}1 & -1\\ -1 & 1\end{pmatrix}$$

Next, we figure out what Theorem 3 tells us about the distribution of F under $H_{0\mathrm{N}}$. Because $m = 1$ here, we have

$$F \xrightarrow{\mathrm{d}} C_1\chi_1^2, \text{ where } C_1 = \frac{CVC^\top}{\bar{S}CP^{-1}C^\top} = \frac{S(1,1)/p_1 + S(2,2)/p_2 - S_\tau^2}{S(1,1)/p_2 + S(2,2)/p_1}$$

which we get by working out the numerator and denominator of the constant:

$$\begin{aligned}
CVC^\top =&(1,\ -1)\begin{pmatrix} \frac{1-p_1}{p_1}S(1,1) & -S(1,2) \\ -S(1,2) & \frac{1-p_2}{p_2}S(2,2)\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}\\
=&\frac{1-p_1}{p_1}S(1,1)+\frac{1-p_2}{p_2}S(2,2)+2S(1,2)\\
=&\frac{S(1,1)}{p_1}+\frac{S(2,2)}{p_2}-S_\tau^2\\
\bar{S}CP^{-1}C^\top =&\bar{S}\cdot(1,\ -1)\cdot\text{diag}\,(\frac{1}{p_1},\frac{1}{p_2})\begin{pmatrix}1\\-1\end{pmatrix}=\{p_1S(1,1)+p_2S(2,2)\}(\frac{1}{p_1}+\frac{1}{p_2})\\
=&\frac{S(1,1)}{p_2}+\frac{S(2,2)}{p_1}
\end{aligned}$$

where for $CVC^\top$ we note $S_\tau^2 = S(1,1)+S(2,2)-2S(1,2)$, as the sample variance of $\{Y_i(1)-Y_i(2) : i=1,\ldots,N\}$. Since $C_1 > 1$ is possible, and recalling from Theorem 3 that $F_\pi|W \xrightarrow{\text{d}} \chi^2_1$, we have that the F statistic derived from a linear models framework may fail even in the case of treatment-control. However, $C_1 \leq 1$ is guaranteed under a balanced design. To see directly that F is proper under a balanced design, note that Lemma 5 applies to give that $F = B$, which we argued above was proper.

As for the claims about the statistic $|\hat{\tau}|$, we first derive Neyman's variance formula found in (B.5). By Proposition 3, $N^{1/2}\hat{\tau} = N^{1/2}C(\hat{\bar{Y}} - \bar{Y}) \xrightarrow{\text{d}} \mathcal{N}(0, CVC^\top)$, which is (B.5), since we have already calculated $CVC^\top$ above.

We now find the asymptotic distribution of $N^{1/2}\hat{\tau}_\pi|W$. We cannot just quote the permutation distributions of X^2, B, and F, and be done immediately. All of those statistics involve some sort of standardization. Yet, we draw on a couple of results in those proofs, special cases of which are relevant here. The imputed potential outcomes are $Y_i^*(1) = Y_i^*(2) = Y_i^{\text{obs}}$, i.e. they satisfy the sharp null hypothesis of no treatment effect whatsoever. We now verify that they satisfy Assumption A almost surely: $\bar{Y}_\bullet^{\text{obs}} = \sum_{j=1}^J N_j\hat{\bar{Y}}(j)/N \xrightarrow{\text{as}} \sum_{j=1}^2 p_j\bar{Y}(j)$ by Lemma 3, applicable because the original potential outcomes satisfy Assumption A. Next, from A.1

$$s^* = \sum_{j=1}^2 \frac{N_j-1}{N-1}\hat{S}(j,j) + \frac{N}{N-1}\sum_{j=1}^2\frac{N_j}{N}\{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\text{obs}}\}^2 \xrightarrow{\text{as}} \sum_{j=1}^2 p_jS(j,j) + p_1p_2\tau^2$$

where for the second piece, we recognize $\sum_{j=1}^2 N_j\{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\text{obs}}\}^2/N$ as the variance of a random variable that takes values $\hat{\bar{Y}}(1)$, $\hat{\bar{Y}}(2)$ with probabilities N_1/N and N_2/N, respectively. We finally have, for all sequences of W, that $\lim_{N\to\infty}\max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2 = 0$, from Lemma 4.

Fix a sequence (W) along which $(\bar{Y}_\bullet^{\text{obs}})$ and (s^*) converge. Then

$$N^{1/2}\{\hat{\bar{Y}}_\pi(1) - \hat{\bar{Y}}_\pi(2)\} \xrightarrow{\text{d}} \mathcal{N}\big(0, \frac{s^*}{p_1}+\frac{s^*}{p_2}-(S_\tau^2)^*\big) \xrightarrow{\text{d}} \mathcal{N}\big(0, \frac{S(1,1)}{p_2}+\frac{S(2,2)}{p_1}+\tau^2\big)$$

The "$\stackrel{\text{d}}{\to}$" comes from using (B.5) on the $Y_i^*(1) = Y_i^*(2) = Y_i^{\text{obs}}$, which satisfy $H_{0\text{N}}$. The "$\stackrel{\text{d}}{=}$" comes from $(S_\tau^2)^* = 0$ for the imputed potential outcomes, and the almost sure limit of s^* found above.

We finally argue that $|\hat{\tau}|$ is not proper. Note under $H_{0\text{N}}$ that $N^{1/2}|\hat{\tau}| \leq_{\text{st}} N^{1/2}|\hat{\tau}_\pi|$ if and only if

$$\frac{S(1,1)}{p_1} + \frac{S(2,2)}{p_2} - S_\tau^2 \leq \frac{S(1,1)}{p_2} + \frac{S(2,2)}{p_1}$$

which is not in general true unless $S(1,1) = S(2,2)$ or $p_1 = p_2$. □

We can easily determine which of V_{F} and V_{N} is larger in which situations. Mimicking a calculation in [23], we have

$$\begin{aligned} V_{\text{F}} - V_{\text{N}} =& \Big\{\frac{S(1,1)}{p_2} + \frac{S(2,2)}{p_1} + \tau^2\Big\} - \Big\{\frac{S(1,1)}{p_1} + \frac{S(2,2)}{p_2} - S_\tau^2\Big\} \\ =& \Big(\frac{1}{p_2} - \frac{1}{p_1}\Big)\{S(1,1) - S(2,2)\} + \tau^2 + S_\tau^2 \end{aligned}$$

As stated, when $V_{\text{F}} < V_{\text{N}}$, the FRT with $|\hat{\tau}|$ fails to control type I error when testing $H_{0\text{N}}$. The case $V_{\text{F}} > V_{\text{N}}$ is fine for the purposes of controlling type I error. Strictly speaking, then, it is of no concern for us. Yet, it presents a different conundrum, both interesting in its own right, and helping to provide a more complete picture of the FRT. Thus, we elect to overview it.

If we take the classical approach to testing $H_{0\text{N}}$ by a z-test motivated by (B.5), i.e. derive a test from the asymptotic fact that

$$\frac{\hat{\tau}}{\hat{S}(1,1)/N_1 + \hat{S}(2,2)/N_2} \leq_{\text{st}} \mathcal{N}(0,1)$$

then the p-value we get is less than the FRT p-value, which was designed to test $H_{0\text{F}}$. We call this the "Neyman approach". This opens the possibility to rejecting $H_{0\text{N}}$ yet failing to reject $H_{0\text{F}}$, which is puzzling because $H_{0\text{F}}$ implies $H_{0\text{N}}$. This phenomenon is the central tenet of [23]. Briefly, the issue is that the FRT is less powerful than Neyman's classical test. As τ deviates from zero more, i.e. $H_{0\text{N}}$ is increasingly violated, it is increasingly likely that $V_{\text{F}} > V_{\text{N}}$ from our calculation above. Intuitively, the FRT's diminished power stems from its use of the variance of all the $\{Y_i^{\text{obs}}\}$. Meanwhile, Neyman's approach uses the variance of $\{Y_i^{\text{obs}} : W_i = 1\}$ and $\{Y_i^{\text{obs}} : W_i = 2\}$ separately, disregarding the between-group component.

We also examine the roles of $S(1,1)$, $S(2,2)$, and p_1 in the variance comparison. If either $p_1 = p_2$ or $S(1,1) = S(2,2)$, then neither the variance nor allocation terms contribute to the comparison. Without loss of generality, $S(1,1) \geq S(2,2)$. It is preferable to take $p_1 \geq p_2$ also, to improve estimation precision, i.e. lower V_{N}. But this makes $V_{\text{N}} \leq V_{\text{F}}$, so we are more likely to observe the paradox. If $V_{\text{N}} \ll V_{\text{F}}$, then the FRT with $|\hat{\tau}|$ has little power against $H_{0\text{N}}$ or even $H_{0\text{F}}$. Using X^2 or B, but not F, in the FRT can improve the power (in fact, if $S(1,1) = S(2,2)$ or $p_1 = p_2$, then using X^2, B, or F can enhance detection of $\tau \neq 0$). On the flip side, if p_2 sufficiently exceeds p_1, and $S(1,1) > S(2,2)$, then $V_{\text{N}} > V_{\text{F}}$,

and the FRT with $|\hat{\tau}|$ fails to control type I error. This overall intuition of assigning smaller p_j to treatments with larger $S(j,j)$ is how we show by counterexample that statistics besides X^2 are not proper in general. Balanced designs are overall optimal for treatment-control experiments. If homoscedasticity holds, then balanced designs give $\hat{\tau}$ the lowest variance. Indeed, $S(1,1)/p_1 + S(2,2)/p_2 - S_\tau^2$ is minimized over all $p_1, p_2 \in [0,1]$ such that $p_1 + p_2 = 1$ when $p_1 = 1/2$. This extends to $J > 2$ also. If heteroscedasticity holds, then balanced designs ensure control of type I error.

We close this section with a short summary that we hope resolves the paradox. Neyman designed an asymptotically conservative test for $H_{0\mathrm{N}}$, while Fisher designed a finite-sample exact test (the FRT) for $H_{0\mathrm{F}}$, both involving $\hat{\tau}$ as a test statistic. Both approaches are of course valid for their stated goals. When both approaches are tried all at once on some set of potential outcomes, we find that Fisher's approach can reject more often than Neyman's. Despite $H_{0\mathrm{F}}$ being stronger than $H_{0\mathrm{N}}$, there is nothing to guarantee that any procedure valid for the former must reject less often than any procedure valid for the latter. Roughly speaking, Fisher's approach, compares $\hat{\tau}$ against $\mathcal{N}(0, V_{\mathrm{F}})$, while Neyman's approach compares it against $\mathcal{N}(0, V_{\mathrm{N}})$. We have shown $V_{\mathrm{F}} - V_{\mathrm{N}}$ might be positive or negative (or zero), and overviewed the situations leading to both possibilities. The moral of the story is, with $\hat{\tau}$, and so long as $V_{\mathrm{F}} \neq V_{\mathrm{N}}$, some audience is destined to be disappointed, depending on whether they care more about type I error control or power. This flaw with using $\hat{\tau}$ gives us another opportunity to recommend studentization, which resolves both issues.

B.4 More on linear models and Huber–White estimation

Linear models are ubiquitous. Here, we give a brief overview of the relevant linear models background for the main text, in particular results that concern hypothesis testing. The linear model assumes

$$y = X\beta + \epsilon \;\leftrightarrow\; \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix} \beta + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

where for N units we observe a response $y_1, \ldots, y_N$ and predictors $x_1, \ldots, x_N \in \mathbb{R}^J$. The predictors are represented compactly in the design matrix $X \in \mathbb{R}^{N \times J}$, which we assume is fixed and full column rank. The J coefficients $\beta \in \mathbb{R}^J$ are fixed and not observed. The noise term ϵ is also not observed, and the only source of randomness. It is standard make one of the following assumptions in linear models: (in order from weakest to strongest)

- $\mathbb{E}\epsilon = 0_N$
- $\mathbb{E}\epsilon = 0_N$ and $\mathrm{Cov}(\epsilon) = \sigma^2 I_N$
- $\epsilon \sim \mathcal{N}(0_N, \sigma^2 I_N)$

Usually, the normality assumption is needed for the classical theory of hypothesis testing with linear models. A general linear hypothesis takes the form $H_0 : C\beta = 0_m$, where $C \in \mathbb{R}^{m\times J}$ is a full row rank contrast matrix, i.e. $C1_J = 0_m$. For pure linear models results, it is not especially important that C is a contrast matrix. In our developments, though, general linear hypotheses with contrast matrices have clearer interpretations. In a discussion of linear models, it is inescapable to encounter some quantities, which we now define. In what follows, let $\mathcal{C}(A)$ be the column space or range of any matrix A. Also, let $H = X(X^\top X)^{-1}X^\top$ be the projection matrix onto $\mathcal{C}(X)$. The ordinary least squares (OLS) estimate of β is

$$\hat{\beta} := \operatorname*{argmin}_{b\in\mathbb{R}^J} |y - Xb|^2 = (X^\top X)^{-1}X^\top y$$

this can be seen multiple ways. One way is by setting the gradient of the objective to zero. A linear algebra approach notices that, at optimum, we must have $y - X\hat{\beta} \perp \mathcal{C}(X)$. The vector $y - X\hat{\beta}$ contains the residuals, and the residual sum of squares is

$$\text{RSS} = |y - X\hat{\beta}|^2 = y^\top(I_N - H)y$$

It helps define the mean squared error (MSE)

$$\hat{\sigma}^2 = \frac{\text{RSS}}{N-J} = \frac{y^\top(I_N - H)y}{N-J}$$

Fundamental results about unbiasedness in linear models are that $\mathbb{E}\hat{\beta} = \beta$ when $\mathbb{E}\epsilon = 0_N$. When we additionally assume $\text{Cov}(\epsilon) = \sigma^2 I_N$, we get $\text{Cov}(\hat{\beta}) = \sigma^2(X^\top X)^{-1}$ and $\mathbb{E}\hat{\sigma}^2 = \sigma^2$. The notation $\hat{\sigma}^2$ is thus explained in that it is a "good" estimator for σ^2. As we do not rely on these results, we omit their proof. An accessible introduction to linear models can be found in [47] or [28]. A more theoretical treatment is given by [18].

We now return to the issue of testing $H_0 : C\beta = 0_m$.

Theorem 6. *Assume the normal linear model $y = X\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0_N, \sigma^2 I_N)$. The F statistic for testing H_0 is*

$$F = \frac{1}{m}(C\hat{\beta})^\top(\hat{\sigma}^2 C(X^\top X)^{-1}C^\top)^{-1}C\hat{\beta}$$

When H_0 is true, we have $F \sim F_{m,N-J}$.

Proof. While we provide a proof, we urge the consultation of [18] for more on this subject. First, write $C = AX$ where $A \in \mathbb{R}^{m\times N}$. Such an A exists because X is full column rank, i.e. $\mathcal{C}(X^\top) = \mathbb{R}^N$. We will show

$$F = \frac{1}{m}(C\hat{\beta})^\top(\hat{\sigma}^2 C(X^\top X)^{-1}C^\top)^{-1}C\hat{\beta} = \frac{y^\top \tilde{H}y/m}{\hat{\sigma}^2} \tag{B.7}$$

where $\tilde{H}$ is the projection matrix onto $\mathcal{C}(HA^\top)$. To that end, note $AH = AXX^+ = CX^+$, so

$$\begin{aligned} y^\top\tilde{H}y =& y^\top(AH)^\top\{(AH)(AH)^\top\}^{-1}(AH)y = (CX^+y)^\top\{AX(X^\top X)^{-1}X^\top A^\top\}^{-1}CX^+y \\ =& (C\hat{\beta})^\top\{C(X^\top X)^{-1}C^\top\}^{-1}C\hat{\beta} \end{aligned}$$

from which (B.7) follows. To prove the null distribution of the F statistic, we make a few more observations. Note the null hypothesis says $0_m = C\beta = AX\beta$, so $\mathbb{E}y = X\beta \in \mathcal{C}(X) \cap \text{Null}(A)$. We next claim

$$\mathcal{C}(X) \cap \text{Null}(A) = \mathcal{C}(X) \cap \text{Null}(AH) = \mathcal{C}(H - \tilde{H})$$

To see the first equality, $z \in \mathcal{C}(X) \cap \text{Null}(A) \leftrightarrow Hz = z$ and $Az = 0_m \leftrightarrow Hz = z$ and $AHz = 0_m \leftrightarrow z \in \mathcal{C}(X) \cap \text{Null}(AH)$. For the second equality, $\text{Null}(AH) = \mathcal{C}(HA^\top)^\perp$, and $\mathcal{C}(HA^\top) \subseteq \mathcal{C}(H) = \mathcal{C}(X)$, so $H - \tilde{H}$ is the projection matrix onto $\mathcal{C}(X) \cap \text{Null}(AH)$. We now have

$$F = \frac{y^\top \tilde{H} y/m}{y^\top (I_N - H)y/(N-J)} = \frac{\epsilon^\top \tilde{H}\epsilon/m}{\epsilon^\top (I_N - H)\epsilon/(N-J)} \sim F_{m,N-J}$$

If H_0 is true, then we have shown $X\beta \in \mathcal{C}(H - \tilde{H})$, so $(H - \tilde{H})X\beta = X\beta \leftrightarrow \tilde{H}X\beta = (H - I_N)X\beta = 0_J$. Thus, under H_0, we have $y^\top \tilde{H} y = \epsilon^\top \tilde{H}\epsilon$. We also have $y^\top (I_N - H)y = \epsilon^\top (I_N - H)\epsilon$ because the full model is true. Since $\epsilon \sim \mathcal{N}(0_N, \sigma^2 I_N)$, we have that, for any projection matrix M, $\epsilon^\top M\epsilon \sim \sigma^2 \chi^2_r$, where r is the rank of M (from a deterministic version of Lemma 1(iii)). The fact that $\mathcal{C}(\tilde{H}) = \mathcal{C}(HA^\top) \subseteq \mathcal{C}(H) \perp \mathcal{C}(I_N - H)$ implies the numerator and denominator are independent (again exploiting normality of ϵ), giving us the claimed null distribution.

While we have already shown everything in the theorem, we draw a parallel with a more familiar form of the F statistic

$$F = \frac{(\text{RSS}_0 - \text{RSS})/(J - r_0)}{\hat{\sigma}^2} \tag{B.8}$$

where we want to test a reduced model $\mathbb{E}y = X\beta \in \mathcal{C}(X_0) \subseteq \mathcal{C}(X)$, RSS_0 is the residual sum of squares from this reduced model, and r_0 is the rank of X_0. We readily deduce that this F statistic matches ours. In our case, the reduced model is $\mathbb{E}y \in \mathcal{C}(H - \tilde{H})$, with rank $r_0 = J - m$ and $\text{RSS}_0 = y^\top (I_N - H + \tilde{H})y$. We also have $\text{RSS} = y^\top (I_N - H)y$, so $\text{RSS}_0 - \text{RSS} = y^\top \tilde{H} y$. □

Because the F statistic stated in Theorem 6 involves standardization by the inverse of an estimated covariance, i.e. $\hat{\sigma}^2 C(X^\top X)^{-1} C^\top$, F is a Wald-type statistic, and thus resembles X^2.

We now tie the linear model into our J-treatment randomized experiment with potential outcomes:

$$Y_i^{\text{obs}} = \bar{Y}(W_i) + \epsilon_i, \text{ for } i = 1, \ldots, N$$

Assume without loss of generality that the first N_1 of the W_i's are 1, the next N_2 of the W_i's are 2, and so on (if this is not so, then permute the observations). Then in matrix form we have

$$Y_i^{\text{obs}} = X\bar{Y} + \epsilon, \text{ where } X = \text{diag}(1_{N_1}, \ldots, 1_{N_J})$$

Notation-wise, we depart from the main text in that we denote the design matrix by X for simplicity, rather than $\mathcal{X}$. The potential outcomes model does not square with the linear model. For instance, due to the CRE treatment assignment mechanism, the ϵ_i are not independent or even uncorrelated. Nevertheless, as mentioned in the main text, we want the

F statistic the linear models framework provides, and to analyze its behavior in the FRT. In so doing, we have seen the most damaging way they collide is that the linear model assumes homoscedasticity. Note that the unknown vector of means $\bar{Y}$ functions as β, and $\hat{\bar{Y}}$ functions as $\hat{\beta}$. One way to see this is to solve the normal equations $X^\top X\hat{\bar{Y}} = X^\top Y^{\text{obs}}$. Then we can see that the F statistic in (3.3) is the same as that in Theorem 6. It is also informative to work out the RSS in the potential outcomes setting:

$$\text{RSS} = \sum_{i=1}^{N}\{Y_i^{\text{obs}} - \hat{\bar{Y}}(W_i)\}^2 = \sum_{i=1}^{N}\sum_{j=1}^{J} W_i(j)\{Y_i^{\text{obs}} - \hat{\bar{Y}}(j)\}^2 = \sum_{j=1}^{J}(N_j - 1)\hat{S}(j,j) \tag{B.9}$$

Proof. (Outline of proof of Theorem 3) Note

$$X^\top X = \text{diag}(N_1, \ldots, N_J), \text{ hence } \lim_{N\to\infty} N(X^\top X)^{-1} = P^{-1}$$

Because each $\hat{S}(j,j) \xrightarrow{P} S(j,j)$, we have $\hat{\sigma}^2 = \text{RSS}/(N-J) \xrightarrow{P} \bar{S}$ by (B.9). By the same argument in Theorem 1, the asymptotic distribution of $m \cdot F$ under $H_{0\text{N}}(C, 0_m)$ follows immediately. For the permutation distribution, the imputed potential outcomes satisfy the sharp null hypothesis (2.2), so they also satisfy $H_{0\text{N}}(C, 0_m)$, hence

$$m \cdot F_\pi | W \xrightarrow{\text{d}} \sum_{j=1}^{m} \lambda_j \big(CV_\pi C^\top (s^* CP^{-1}C^\top)^{-1}\big)\xi_j^2$$

Because the analog for $\bar{S}$ for imputed potential outcomes is $\sum_{j=1}^{J} p_j S^*(j,j) = s^*$, we have

$$CV_\pi C^\top = s^* C(P^{-1} - 1_J 1_J^\top)C^\top = s^* CP^{-1}C^\top$$

hence the eigenvalue weight of each ξ_j^2 is 1. □

One reason the F statistic fails to be proper is that comes from the linear models framework and not the potential outcomes one. An example of an incorrect conclusion we can draw is $\text{Cov}(\hat{\bar{Y}}) = \sigma^2(X^\top X)^{-1}$, i.e. $\text{Var}(\hat{\bar{Y}}(j)) = \sigma^2/N_j$ for $j = 1, \ldots, J$. Recall the true covariance structure of $\hat{\bar{Y}}$ is given in Proposition 3.

We provide some more background on Huber–White estimation. It is intended to estimate $\text{Cov}(\hat{\beta})$ in a robust way when the linear model is possibly misspecified, as it is in the potential outcomes framework. For more details on Huber–White estimation, see [2]. Say $(x, y) \in \mathbb{R}^J \times \mathbb{R}$ is a random variable. If $\beta = (\mathbb{E}xx^\top)^{-1}\mathbb{E}yx$, then $x^\top\beta$ is the best linear predictor of y given x. We do not account for the intercept term. If we define $e = y - x^\top\beta$, then $\mathbb{E}ex = 0_J$. If we observe N iid samples $(x_1, y_1), \ldots, (x_N, y_N)$ of (x, y), put

$$X = \begin{pmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{pmatrix},\ y = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix},\ \hat{\beta} = X^+ Y,\ \epsilon = Y - X\beta,\ \hat{\epsilon} = Y - X\hat{\beta}$$

in particular $\epsilon_i = y_i - x_i^\top \beta$, $\hat{\epsilon}_i = y_i - x_i^\top \hat{\beta}$. We now provide a theorem for the asymptotic distribution of $\hat{\beta}$ along with a rough justification. The reason we switch from fixed design matrix X to random is that now the sample size $N \to \infty$, rather than being fixed as it was at the start of our linear models discussion.

Theorem 7. *Under some regularity on the joint distribution of (x, y), for instance existence of its covariance matrix, we have*

$$N^{1/2}(\hat{\beta} - \beta) \xrightarrow{\text{d}} \mathcal{N}\big(0_J, (\mathbb{E}xx^\top)^{-1}\mathbb{E}e^2xx^\top(\mathbb{E}xx^\top)^{-1}\big)$$

The covariance matrix above is estimated by

$$\hat{D}_{\text{HW}} = N(X^\top X)^{-1}X^\top \operatorname{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_N^2)X(X^\top X)^{-1}$$

or, more compactly, $NX^+ \operatorname{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_N^2)X^{+\top}$.

Proof. As stated, we only present the main ideas, brushing aside the messier technical details and regularity conditions. Note

$$0_J = X^\top \hat{\epsilon} = X^\top y - X^\top X\hat{\beta} = X^\top(X\beta + \epsilon) - X^\top X\hat{\beta}$$

so $X^\top X(\hat{\beta} - \beta) = X^\top \epsilon$, which gives

$$\hat{\beta} - \beta = X^+\epsilon = \big(\frac{1}{N}\sum_{i=1}^N x_i x_i^\top\big)^{-1}\big(\frac{1}{N}\sum_{i=1}^N e_i x_i\big)$$

where we have multiplied and divided by $1/N$. Now by the weak law of large numbers and central limit theorem respectively (here we need some regularity on the joint distribution of (x, y), and the iid assumption)

$$\frac{1}{N}\sum_{i=1}^N x_i x_i^\top \xrightarrow{P} \mathbb{E}xx^\top, \ \frac{1}{N^{1/2}}\sum_{i=1}^N e_i x_i \xrightarrow{\text{d}} \mathcal{N}(0_J, \mathbb{E}e^2xx^\top)$$

Hence by Lemma 7, the asymptotic distribution of $N^{1/2}(\hat{\beta} - \beta)$ follows. Using the weak law again

$$\begin{aligned}(\mathbb{E}xx^\top)^{-1}\mathbb{E}e^2xx^\top(\mathbb{E}xx^\top)^{-1} &\approx \big(\frac{1}{N}\sum_{i=1}^N x_i x_i^\top\big)^{-1}\big(\frac{1}{N}\sum_{i=1}^N \hat{\epsilon}_i^2 x_i x_i^\top\big)\big(\frac{1}{N}\sum_{i=1}^N x_i x_i^\top\big)^{-1} \\ &= N(X^\top X)^{-1}X^\top \operatorname{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_N^2)X(X^\top X)^{-1}\end{aligned}$$

which motivates the estimator $\hat{D}_{\text{HW}}$. □

We now work out what $\hat{D}_{\text{HW}}$ is in the potential outcomes framework. [89] made similar calculations in the special case of treatment-control. In our case, $y = Y^{\text{obs}}$, $X =$

$\mathrm{diag}(1_{N_1}, \ldots, 1_{N_J})$, $\beta = \bar{Y}$, and $\hat{\beta} = \hat{\bar{Y}}$, so $\hat{\epsilon}_i = Y_i^{\mathrm{obs}} - \hat{\bar{Y}}(W_i)$. The Moore-Penrose pseudo inverse is

$$X^+ = (X^\top X)^{-1} X^\top = \mathrm{diag}\left(\frac{1}{N_1}, \ldots, \frac{1}{N_J}\right) \mathrm{diag}(1_{N_1}^\top, \ldots, 1_{N_J}^\top) = \mathrm{diag}(1_{N_1}^\top/N_1, \ldots, 1_{N_J}^\top/N_J)$$

which leads to

$$\hat{D}_{\mathrm{HW}} = N \cdot \mathrm{diag}(1_{N_1}^\top/N_1, \ldots, 1_{N_J}^\top/N_J)\, \mathrm{diag}(\hat{\epsilon}_1^2, \ldots, \hat{\epsilon}_N^2)\, \mathrm{diag}(1_{N_1}^\top/N_1, \ldots, 1_{N_J}^\top/N_J)^\top$$

This is a multiplication of (block) diagonal matrices. We relabel $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_N$ as $\hat{\epsilon}_{1,1}, \ldots, \hat{\epsilon}_{1,N_1}$, $\ldots, \hat{\epsilon}_{J,1}, \ldots, \hat{\epsilon}_{J,N_J}$ for convenience. In other words, the subscripts also specify which of the J groups the i-th observation belongs to. Thus, $\hat{D}_{\mathrm{HW}}$ is a diagonal matrix with (j,j)-entry

$$\frac{N}{N_j^2} 1_{N_j}^\top \mathrm{diag}(\hat{\epsilon}_{j,1}^2, \ldots, \hat{\epsilon}_{j,N_j}^2) 1_{N_j} = \frac{N}{N_j^2} \sum_{i=1}^N W_i(j) \{Y_i^{\mathrm{obs}} - \hat{\bar{Y}}(j)\}^2 = \frac{N(N_j - 1)}{N_j^2} \hat{S}(j,j)$$

This converges in probability to $S(j,j)/p_j$, so $\hat{D}_{\mathrm{HW}} \xrightarrow{P} D$. Thus, if we use the Wald-statistic along with the Huber-White variance estimator, we get

$$X^2_{\mathrm{HW}} = N(C\hat{\bar{Y}})^\top (C \hat{D}^2_{\mathrm{HW}} C^\top)^{-1} C\hat{\bar{Y}}$$

which has the same asymptotic properties as X^2, so it is proper. Note $\hat{D}^2_{\mathrm{HW}} \prec \hat{D}$, so $X^2_{\mathrm{HW}} > X^2$. In general, Huber–White tends to underestimate the true variance, and by a possibly serious amount when N is small. This is problematic, e.g., when confidence intervals coming from Huber–White tend to be narrower than they should be. Finite-sample improvements on the Huber–White estimator have been widely studied [64]. As this takes us outside the realm of the main text, we omit any discussion of the matter.

B.5 More on the one-way layout

Recall that the classical one-way layout or ANOVA hypothesis is $H_{0\mathrm{N}} : \bar{Y}(1) = \cdots \bar{Y}(J)$. This is (2.1) with $C = (1_{J-1}, -I_{J-1})$ and $x = 0_{J-1}$, though other choices of C with $\mathcal{C}(C)^\perp = \mathcal{C}(1_J)$ work just as well. We have already motivated the linear models origins of the classical F statistic (3.3). We have also shown in Proposition 6 how the more common, alternative form (4.1) equates to (3.3). Nevertheless, we would like to motivate (4.1) directly. First, we recall $\bar{Y}_\bullet^{\mathrm{obs}} = \sum_{i=1}^N Y_i^{\mathrm{obs}}$ and define $s_\bullet^{\mathrm{obs}} = \sum_{i=1}^N (Y_i^{\mathrm{obs}} - \bar{Y}_\bullet^{\mathrm{obs}})^2/(N-1)$. Also recall the RSS is $\sum_{i=1}^N \sum_{j=1}^J W_i(j) \{Y_i^{\mathrm{obs}} - \hat{\bar{Y}}(j)\}^2$ given (B.9). We also define the total sum of squares (TSS) and model/treatment sum of squares (MSS)

$$\mathrm{TSS} = \sum_{i=1}^N (Y_i^{\mathrm{obs}} - \bar{Y}_\bullet^{\mathrm{obs}})^2 = (N-1) s_\bullet^{\mathrm{obs}}, \ \mathrm{MSS} = \sum_{i=1}^N \{\hat{\bar{Y}}(W_i) - \bar{Y}_\bullet^{\mathrm{obs}}\}^2 = \sum_{j=1}^J N_j \{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\mathrm{obs}}\}^2$$

The idea of MSS is that $X\hat{\bar{Y}} = HY^{\mathrm{obs}}$ is the vector of fitted or predicted values, which has i-th entry $\hat{\bar{Y}}(W_i)$. An important ANOVA identity is the decomposition of sum of squares

$\mathrm{TSS} = \mathrm{MSS} + \mathrm{RSS}$. This is easily seen with quadratic forms. Let H_1 be the projection matrix onto $\mathcal{C}(1_N)$ and recall $H = X(X^\top X)^{-1}X^\top$. Then

$$\mathrm{TSS} = (Y^{\mathrm{obs}})^\top(I_N - H_1)Y^{\mathrm{obs}} = (Y^{\mathrm{obs}})^\top(H - H_1)Y^{\mathrm{obs}} + (Y^{\mathrm{obs}})^\top(I_N - H)Y^{\mathrm{obs}} = \mathrm{MSS} + \mathrm{RSS}$$

Taking the reduced versus full model approach to the F-test in (B.8), we have

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS})/(J - r_0)}{\hat{\sigma}^2} = \frac{(\mathrm{TSS} - \mathrm{RSS})/(J-1)}{\hat{\sigma}^2} = \frac{\mathrm{MSS}/(J-1)}{\hat{\sigma}^2}$$

which matches (4.1). Note that here we are testing $\mathbb{E}y \in \mathcal{C}(1_N)$ from a linear models point of view. This particular reduced model with only an intercept is often called the "null model", not to be confused with a reduced model $\mathbb{E}y \in \mathcal{C}(X_0)$ associated with a general null hypothesis. The RSS under the null model is the TSS.

In [24], it was shown that F asymptotically follows an $F_{J-1,N-J}$ (or equivalently a $\chi^2_{J-1}/(J-1)$) distribution under the sharp null $H_{0\mathrm{F}} : Y_i(1) = \cdots = Y_i(j)$ for $i = 1, \ldots, N$. They thus established a potential outcomes counterpart of the classic result from iid samples. But the asymptotic permutation distribution of $F_\pi|W$ does not stochastically dominate the randomization distribution of F under $H_{0\mathrm{N}}$. This was both argued heuristically by looking at the expectation of the numerator and denominator of the F statistic and via a simulation. They proposed (4.2) as a fix. We have shown through Proposition 6 that (4.2) matches (3.1). For computation purposes, in the ANOVA setting, we recommend the form (4.2) because it avoids computing a quadratic form with an inverse matrix. This again emphasizes that we should take advantage of special cases such as ANOVA or treatment-control to get (3.1) in as simple a form as possible.

B.6 More on the Box-type and Wald-type statistics

The Box-type and studentized (or Wald-type) statistic have also been studied effusively by [15, 76, 54, 34, 35].

Instead of (3.1), the studentized statistic appearing in [76, 54] for the special case $H_{0\mathrm{N}}(C, 0_m)$ is $N(M\hat{\bar{Y}})^\top(M\hat{D}M)^- M\hat{\bar{Y}}$, where we recall $M = C^\top(CC^\top)^{-1}C$ is the projection matrix onto $\mathcal{C}(C^\top)$. We modify this statistic to accommodate nonzero x:

$$X_B^2 = N(\hat{\bar{Y}} - C^+x)^\top M(M\hat{D}M)^+ M(\hat{\bar{Y}} - C^+x). \tag{B.10}$$

In fact, any choice of $z \in \mathbb{R}^J$ such that $Cz = x$ may be used in place of C^+x. The next result shows it is equivalent to use X^2 or X_B^2.

Lemma 12. *We have $X^2 = X_B^2$. Hence, X^2 depends on C only through $\mathcal{C}(C^\top)$. As a consequence, we may take C to be full row rank without loss of generality.*

Proof. For full generality, we consider $C \in \mathbb{R}^{m\times J}$ without assuming it is full row rank, but we must additionally assume $x \in \mathcal{C}(C)$, since $\mathcal{C}(C) \neq \mathbb{R}^m$ is now possible. Then we must modify the test statistic (3.1) into $X^2 = N(C\hat{\bar{Y}} - x)^\top(C\hat{D}C^\top)^-(C\hat{\bar{Y}} - x)$. To be clear, we

use the generalized inverse of $C\hat{D}C^\top$ because we cannot be sure it is invertible. To equate this to (B.10), it is enough to show

$$M(M\hat{D}M)^+M = C^\top(C\hat{D}C^\top)^-C$$

By verifying the definition of generalized inverse, which merely states that a matrix G is a generalized inverse of a matrix A if $AGA = A$, we have

$$(M\hat{D}M)^- = C^\top(C\hat{D}C^\top)^-C$$

which holds no matter which generalized inverses we pick. Since M is the projection matrix onto $\mathcal{C}(C^\top)$, we have $CM = C$ and $MC^\top = C^\top$, so

$$M(M\hat{D}M)^-M = C^\top(C\hat{D}C^\top)^-C$$

Finally, because the Moore-Penrose pseudoinverse is a particular generalized inverse, it remains to argue that the LHS is invariant to the choice of generalized inverse. Because $\hat{D}$ is invertible, this happens if $\hat{D}^{1/2}M(M\hat{D}M)^-M\hat{D}^{1/2}$ is invariant in the same way, which is true because it is the (unique) projection matrix for $\hat{D}^{1/2}M$ [39]. □

Thus, the advantage of viewing X^2 using (B.10) is to see that it is invariant to the choice of C provided the row space is unchanged. Using a full row rank C is not just for mathematical convenience. Removing redundant rows from C beforehand also reduces the dimension of the problem. If C is full row rank, then we must have $m \leq J-1$ because C is also a contrast matrix.

The FRT with X^2 is robust for two null hypotheses. It asymptotically controls type I error for $H_{0\mathrm{N}}(C,x)$, while retaining finite-sample exactness for the sharp null (2.2). In particular, it is also robust to treatment effect heterogeneity. An implication is that, if we do not believe we are in an asymptotic regime and do not feel comfortable reaching a conclusion on the weak null, we can still walk away with a verdict on the sharp null. As another heuristic, we are always at liberty to sample iid from the distribution $X^2_\pi|W$. Thus, we can check how close it is to its asymptotic distribution of χ^2_m. The test statistic X^2 is also intuitive. It is roughly a norm for $C\hat{\bar{Y}} - x$, which should be close to zero when $H_{0\mathrm{N}}(C,x)$ is true, and large otherwise. This motivates looking at the right tail to compute a p-value in the FRT. On the other hand, the imputed potential outcomes were designed to make $C\hat{\bar{Y}}_\pi - x$ close to zero (as $C\bar{Y}^* - x$ is exactly zero). Thus, the values of X^2_π have no tendency to get larger when $H_{0\mathrm{N}}(C,x)$ is violated. This explains heuristically the power of the FRT with X^2. [19] and [76] further discuss the power of permutation tests.

As mentioned, the FRT with any statistic T is finite sample exact for testing the sharp null hypothesis (2.2). From the perspective of testing $H_{0\mathrm{N}}(C,x)$, however, $\tilde{x}$ constitutes a nuisance parameter. This is the case even with a proper statistic T. Theorem 1 shows that, asymptotically, the choice of $\tilde{x}$ does not matter. Looking at its proof, this depends on $(\bar{Y})_{N\geq 2J}$ being bounded above in norm. In finite samples, however, varying $\tilde{x}$ can result in different p-values being obtained because of a different sharp null being tested. If the experimenter only cares about testing $H_{0\mathrm{N}}(C,x)$ and wants a test free from the effects of $\tilde{x}$,

then a supremum of p-values over a large grid of possible $\tilde{x}$ values should be used, following [25, 9]. Alternatively, use a χ^2-approximation to determine the p-value. Despite s^* being known to us, it should not be used in the calculation of T_π since it must emulate the form of T, which only involves sample quantities. In other words, s^* is known because the imputed potential outcomes have very special structure, being strictly additive. We cannot (and do not) make this assumption on the original potential outcomes.

We now shift gears from the Wald-type to the Box-type statistic B in (3.2), embarking by motivating its form. This statistic was derived in [15] by performing a Box-type approximation [12] of matching the first two moments of the quadratic form $\hat{\bar{Y}}^\top M \hat{\bar{Y}}$ with $g \cdot \chi^2_f$ on the quadratic form. This led to $N\hat{\bar{Y}}^\top M\hat{\bar{Y}}/\operatorname{tr}(MD) \overset{d}{\approx} \chi^2_f/f$ for some f. They figured out the exact distribution of B in an ANOVA model with heteroscedastic normal errors, but for convenience sake used an F-distribution as an approximation. In simulations, they found this statistic to have strong empirical small-sample performance. A computational advantage of B is that, unlike X^2, it does not involve a matrix inverse.

To derive the statistic B, we assume a heteroscedastic linear model. For $j = 1, \ldots, J$, $\{Y_i^{\text{obs}} : W_i = j\}$ are an iid sample from $\mathcal{N}(\bar{Y}(j), S(j,j))$. The Y_i^{obs} are also independent for distinct treatments j. Had we also assumed $S(1,1) = \cdots = S(J,J)$, then this is the one-way ANOVA model. Under the model of [15], we have $\hat{\bar{Y}} \sim \mathcal{N}(\bar{Y}, D)$ where we recall D comes from (2.4). As before, with a slight abuse of notation, we let $p_j = N_j/N$ or its limit. We want to test $H_0 : C\bar{Y} = 0_m$, the same hypothesis as (2.1). Under this model and H_0, $X^2 \overset{\text{d}}{\to} \chi^2_m$. As stated in [76], this can lead to a poor finite sample approximation.

To improve this, recall $M = C^\top (CC^\top)^{-1} C$, and start with the quadratic form

$$N\hat{\bar{Y}}^\top M \hat{\bar{Y}} \sim \sum_{j=1}^{J} \lambda_j(MD)\xi_j^2$$

where the distribution is obtained from Lemma 1. Since D is usually unknown, the $\lambda_j(MD)$ must be estimated, which often leads to unsatisfactory approximations. As a prelude to deriving B as an improvement, we take a moment to clarify what χ^2_f means when $f \notin \mathbb{Z}^+$. A random variable X has a gamma distribution, written $X \sim \Gamma(a, b)$, if it has density

$$f(x) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b}, \text{ for } x > 0, \text{ where } \Gamma(a) = \int_0^\infty x^{a-1} e^{-x} \mathrm{d}x$$

is the gamma function. For $f \in \mathbb{Z}^+$, χ^2_f is defined as the distribution of $\sum_{j=1}^f \xi_j^2$. In this case, $\chi^2_f \overset{\text{d}}{=} \Gamma(f/2, 2)$ (this follows by comparing the moment generating functions; we also need to use $\chi^2_1 \overset{\text{d}}{=} \xi^2$ to get its density directly). Since the gamma distribution allows any shape and scale parameters $a, b > 0$, it makes sense to define χ^2_f as $\Gamma(f/2, 2)$ for arbitrary $f > 0$. If $X \sim \chi^2_f$, then $\mathbb{E}X = f$ and $\operatorname{Var}(X) = 2f$.

Lemma 13. *Under $H_0 : C\bar{Y} = 0_m$, we have*

$$B = \frac{N}{\operatorname{tr}(M\hat{D})} \hat{\bar{Y}}^\top M \hat{\bar{Y}} \overset{d}{\approx} F_{\hat{f}, \hat{f}_0}, \text{ where } \hat{f} = \frac{\operatorname{tr}^2(M\hat{D})}{\operatorname{tr}(M\hat{D}M\hat{D})}, \ \hat{f}_0 = \frac{\operatorname{tr}^2(M\hat{D})}{\operatorname{tr}(D_M^2 \hat{D}^2 P_0)}$$

where $D_M = \operatorname{diag}(M)$ *is a diagonal matrix that matches the main diagonal of* M*, and* $P_0 = \operatorname{diag}(N_1 - 1, \ldots, N_J - 1)^{-1}$*.*

Most of the occurrences M can be replaced by D_M, eg $\operatorname{tr}(M\hat{D}) = \operatorname{tr}(D_M\hat{D})$ because $\hat{D}$ is diagonal, and, as far as trace is concerned, only the main diagonal of its argument matters. Unlike [15], we write everything in terms of M as opposed to D_M whenever possible.

Proof. Put

$$f = \frac{\operatorname{tr}^2(MD)}{\operatorname{tr}(MDMD)},\ f_0 = \frac{\operatorname{tr}^2(MD)}{\operatorname{tr}(D_M^2 D^2 P_0)}$$

which are the quantities $\hat{f}$ and $\hat{f}_0$ approximate. Our aim is to show

$$B = \tilde{F}/F_0,\ \text{where } \tilde{F} = \frac{N}{\operatorname{tr}(MD)}\hat{\bar{Y}}^\top M\hat{\bar{Y}} \overset{d}{\approx} \chi^2_f/f,\ F_0 = \frac{\operatorname{tr}(M\hat{D})}{\operatorname{tr}(MD)} \overset{d}{\approx} \chi^2_{f_0}/f_0 \tag{B.11}$$

then from the fact that $\tilde{F}$ and F_0 are independent, we get $F \overset{d}{\approx} F_{f,f_0}$. The independence is because of the normality assumption: $\hat{\bar{Y}}$ and $\hat{D}$ are independent.

The idea to show both approximations in (B.11) is to use a "Box-type" approximation [12]. We can do a Box-type approximation on the numerator $\tilde{F}$ only and leave it at that, but doing it on the denominator F_0 also gives superior empirical performance. We approximate $N\bar{X}^\top M\bar{X} \overset{d}{\approx} g\chi^2_f$, where g and f are chosen to make the first 2 moments match. Recalling $N\hat{\bar{Y}}^\top M\hat{\bar{Y}} \sim \sum_{j=1}^J \lambda_j(MD)\xi_j^2$, the mean and variance are

$$\begin{aligned}\mathbb{E}(N\hat{\bar{Y}}^\top M\hat{\bar{Y}}) =& \sum_{j=1}^J \lambda_j(MD) = \operatorname{tr}(MD)\\ \operatorname{Var}(N\hat{\bar{Y}}^\top M\hat{\bar{Y}}) =& 2\sum_{j=1}^J \lambda_j^2(MD) = \operatorname{tr}(MDMD)\end{aligned}$$

For the variance, we have used for diagonalizable A that the eigenvalues of A^2 are the squared eigenvalues of A. To equate the corresponding moments of $g\chi^2_f$, we need $gf = \operatorname{tr}(MD)$, and $2g^2 f = 2\operatorname{tr}(MDMD)$. After some algebra, we get

$$\tilde{F} = \frac{N\hat{\bar{Y}}^\top M\hat{\bar{Y}}}{\operatorname{tr}(MD)} \overset{d}{\approx} \chi^2_f/f$$

which is the first part of (B.11). For the second part, note

$$\operatorname{tr}(M\hat{D}) = N\sum_{j=1}^J \frac{m_{jj}\hat{S}(j,j)}{N_j} \overset{\text{d}}{=} N\sum_{j=1}^J \frac{m_{jj}S(j,j)}{N_j(N_j-1)}V_j$$

where $V_j \sim \chi^2_{N_j-1}$ and are independent. This follows from the normal distribution theory: $\hat{S}(j,j) \sim S(j,j)\chi^2_{N_j-1}/(N_j-1)$. Hence

$$\begin{aligned}\mathbb{E}\,\mathrm{tr}(M\hat{D}) =& \mathrm{tr}(MD)\\ \mathrm{Var}\left(\mathrm{tr}(M\hat{D})\right) =& 2N^2\sum_{j=1}^{J}\frac{m_{jj}^2 S^2(j,j)}{N_j^2(N_j-1)} = 2\,\mathrm{tr}(D_M^2 D^2 P_0)\end{aligned}$$

To equate the moments of $\mathrm{tr}(M\hat{D})$ with $g_0\chi^2_{f_0}/f_0$, we need $g_0 = \mathrm{tr}(MD)$, and $2g_0^2/f_0 = 2\,\mathrm{tr}(D_M^2 D^2 P_0)$. Some algebra now shows the second part of (B.11). □

We have seen in the main text that B is not proper for the FRT in a potential outcomes setting. It is logical that B would suffer a similar drawback in the iid samples setting discussed here. Its distribution is not pivotal, and so it should not be used in a permutation test. Here is another informative result about the behavior of B.

Corollary 6. *In the setting of Theorem 2,*

$$B \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} a_j\xi_j^2 \text{ and } B_\pi|W \xrightarrow{\mathrm{d}} \sum_{j=1}^{m} b_j\xi_j^2, \text{ where each } a_j, b_j \geq 0, \sum_{j=1}^{m} a_j \leq 1, \text{ and } \sum_{j=1}^{m} b_j = 1$$

In particular, $\mathbb{E}B \leq 1$ *and* $\mathbb{E}(B_\pi|W) = 1$.

Proof. Throughout, we make repeated use of Lemma 8 and (A.4). Since trace is a sum of eigenvalues, it is clear that

$$\sum_{j=1}^{m}\lambda_j(MP^{-1})/\,\mathrm{tr}(MP^{-1}) = 1$$

Also, each $\lambda_j(MP^{-1}) = \lambda_j(MP^{-1}M) \geq 0$ because $P \succeq 0$ (Note we consider MDM because the product of two symmetric matrices is not symmetric in general). For the situation under $H_{0\mathrm{N}}(C, 0_m)$, each $\lambda_j(MV) \leq \lambda_j(MD)$, which gives

$$\sum_{j=1}^{m}\lambda_j(MV)/\,\mathrm{tr}(MD) \leq \sum_{j=1}^{m}\lambda_j(MD)/\,\mathrm{tr}(MD) = 1$$

and each $\lambda_j(MD) \geq 0$. □

This corollary shows the Box-type statistic B is somewhat promising (e.g., there is no analog for F). Still, it is not able to control type I error, i.e. it does not imply the criterion of Proposition 4. Indeed, $\sum_{j=1}^m a_j \leq \sum_{j=1}^m b_j$ does not guarantee $\sum_{j=1}^m a_j\xi_j^2 \leq_{st} \sum_{j=1}^m b_j\xi_j^2$.

Corollary 1 states that B is proper under homoscedasticity. In light of this, before using Brunner's statistic, it could be sensible to make sure $\{\hat{S}(1,1), \ldots, \hat{S}(J,J)\}$ are close to each other. The same comments apply to the F statistic.

B.7 More on vector potential outcomes

[58] also provide tools to handle vector potential outcomes: a central limit theorem and a law of large numbers. We need to generalize Propositions 2 and 3. Recall that these came from Proposition 3 and Theorem 5 in [58]. In actuality, they allowed for vector potential outcomes. The next result was stated in proving Theorem 5, we now flesh out the necessary calculations to arrive at it.

Proposition 8. *Under Assumption D, we have* $\hat{\bar{Y}} \xrightarrow{P} \bar{Y}$ *and* $\hat{S}(j,j) \xrightarrow{P} S(j,j)$ *for* $j = 1,\dots,J$. *If Assumption D and* $H_{0\mathrm{N}}(C,x)$ *holds for all* $N \geq J(d+1)$, *then* $N^{1/2}(C\hat{\bar{Y}} - x) \xrightarrow{\mathrm{d}} \mathcal{N}(0_m, CVC^\top)$, *where*

$$V = \lim_{N\to\infty} N \cdot \mathrm{Cov}(\hat{\bar{Y}}) = \lim_{N\to\infty} \begin{pmatrix} \frac{N-N_1}{N_1}S(1,1) & -S(1,2) & \cdots & -S(1,J) \\ -S(2,1) & \frac{N-N_2}{N_2}S(2,2) & \cdots & -S(2,J) \\ \vdots & \vdots & \ddots & \vdots \\ -S(J,1) & -S(J,2) & \cdots & \frac{N-N_J}{N_J}S(J,J) \end{pmatrix}$$

Note the formula for V is identical to (2.3). However, we emphasize that the $S(j,k)$ are now themselves matrices. Theorem 5 in [58] writes $C\bar{Y} = \sum_{j=1}^{J} A_j \bar{Y}(j)$ for matrices $A_1,\dots,A_J \in \mathbb{R}^{m\times d}$.

Proof. We only need to compute $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$. This is done in [25]. We do it also, for completeness. It is enough to show

$$\mathrm{Cov}(\hat{\bar{Y}}(1)) = \frac{N-N_1}{N_1 N} S(1,1),\ \mathrm{Cov}(\hat{\bar{Y}}(1), \hat{\bar{Y}}(2)) = -S(1,2)/N \tag{B.12}$$

We generalize (B.1). Let $X \in \mathbb{R}^{N\times J}$, $Y \in \mathbb{R}^{N\times K}$ be data matrices with (i,j)-entries x_{ij},y_{ij}, which are the i-th observation of the j-th variable in X and Y, respectively. Recall $V_N = I_N - 1_N 1_N^\top/N$. Then $X^\top V_N Y$ has $(N-1)$ times the sample covariance of X_j and Y_k as its (j,k)-entry, where X_j is the j-th column of X, and Y_k is the k-th column of Y. Hence if $Y(j) \in \mathbb{R}^{N\times d}$ has i-th row $Y_i(j)^\top$, then $S(1,1) = Y(1)^\top V_N Y(1)/(N-1)$ and $S(1,2) = Y(1)^\top V_N Y(2)/(N-1)$. Now we can show (B.12). The idea is the same as proving part 2 of Lemma 9, with vector instead of scalar responses. The first part is

$$\begin{aligned} \mathrm{Cov}(\hat{\bar{Y}}(j)) =& \mathrm{Cov}\left(\frac{1}{N_j}\sum_{i=1}^{N} W_i(j) Y_i(j)\right) = \frac{1}{N_j^2} Y(j)^\top \mathrm{Cov} \begin{pmatrix} W_1(j) \\ \vdots \\ W_N(j) \end{pmatrix} Y(j) \\ =& \frac{1}{N_j^2} \frac{N_j(N-N_j)}{N(N-1)} Y(j)^\top V_N Y(j) = \frac{N-N_j}{N_j N} S(j,j) \end{aligned}$$

and the second part is

$$\begin{aligned}\mathrm{Cov}(\hat{\bar{Y}}(j),\hat{\bar{Y}}(k)) =& \mathrm{Cov}\left(\frac{1}{N_j}\sum_{i=1}^{N} W_i(j)Y_i(j), \frac{1}{N_k}\sum_{i=1}^{N} W_i(k)Y_i(k)\right)\\ =& \frac{1}{N_jN_k} Y(j)^\top \mathrm{Cov}\left(\begin{pmatrix} W_1(j)\\ \vdots \\ W_N(j)\end{pmatrix}, \begin{pmatrix} W_1(k)\\ \vdots \\ W_N(k)\end{pmatrix}\right) Y(k)\\ =& \frac{1}{N_jN_k}\frac{-N_jN_k}{N(N-1)} Y(j)^\top V_N Y(k) = -S(j,k)/N\end{aligned}$$

where Lemma 9 part 1 is used in the second equality of both calculations. □

Recall $D = \mathrm{diag}\{S(1,1)/N_1, \ldots, S(J,J)/p_J\}$ and $\hat{D} = \mathrm{diag}\{\hat{S}(1,1)/p_1, \ldots, \hat{S}(J,J)/p_J\}$, where $\hat{D}$ estimates D, i.e., (2.4) remains true for vector potential outcomes. The only thing to be mindful of is that these are now block diagonal matrices. It is still true that $V = D - S \preceq D$. Thus, $\hat{D}$ is an asymptotically conservative estimator for $N \cdot \mathrm{Cov}(\hat{\bar{Y}})$ in the sense that $\lim_{N\to\infty} N \cdot \mathrm{Cov}(\hat{\bar{Y}}) \preceq \mathrm{plim}_{N\to\infty} \hat{D}$.

B.8 Technical matters for FRT

Assumption B can just as well be stated with an uncentered fourth moment assumption, which we now show. The details are messier, which is why we went with a centered fourth moment in the main text.

Lemma 14. *Let $\{Y_{N,i} : i = 1, \ldots, N\}$, or $\{Y_i\}$ for short, be a sequence of populations indexed by $N \in \mathbb{Z}^+$. If there exists $L < \infty$ such that $\sum_{i=1}^N Y_i^4/N \le L$, for all $N \in \mathbb{Z}^+$, then to each $c \in \mathbb{R}$ there exists $L_c < \infty$ depending on L, c such that $\sum_{i=1}^N (Y_i - c)^4/N \le L_c$. In particular, if (c_N) converges, then the sequence $\left(\sum_{i=1}^N Y_i^4/N\right)$ is bounded if and only if the sequence $\left(\sum_{i=1}^N (Y_i - c_N)^4/N\right)$ is.*

Proof. We have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}(Y_i - c)^4 =& \frac{1}{N}\sum_{i=1}^{N} Y_i^4 - \frac{4}{N}c\sum_{i=1}^{N} Y_i^3 + \frac{6}{N}c^2\sum_{i=1}^{N} Y_i^2 - \frac{4}{N}c^3\sum_{i=1}^{N} Y_i + c^4\\ \le& \frac{1}{N}\sum_{i=1}^{N} Y_i^4 + \frac{4}{N}|c|\sum_{i=1}^{N}|Y_i|^3 + \frac{6}{N}c^2\sum_{i=1}^{N} Y_i^2 + \frac{4}{N}|c|^3\sum_{i=1}^{N}|Y_i| + c^4\\ \le& L + 4|c|L^{3/4} + 6c^2L^{1/2} + 4|c|^3L^{1/4} + c^4 := L_c\end{aligned}$$

where in the 3rd line we use Lyapunov's inequality that, for $0 < k \le 4$

$$\left(\frac{1}{N}\sum_{i=1}^{N}|Y_i|^k\right)^{1/k} \le \left(\frac{1}{N}\sum_{i=1}^{N} Y_i^4\right)^{1/4} = L^{1/4}$$

If (c_N) converges, then it is bounded: there are $a, b \in \mathbb{R}$ such that each $c_n \in [a, b]$. Now $c \mapsto \sum_{i=1}^N (Y_i - c)^4/N$ is a continuous map, so it attains a max on $[a, b]$, hence

$$\frac{1}{N}\sum_{i=1}^N (Y_i - c_N)^4 \leq \max_{c\in[a,b]} \frac{1}{N}\sum_{i=1}^N (Y_i - c)^4 \leq \max_{c\in[a,b]} L_c < \infty$$

To show the converse, we use the same argument with $Y_i \leftarrow Y_i - c_N$ and $c_N \leftarrow -c_N$. □

As stated, we changed $Y_i(j)^4$ to $\{Y_i(j) - \bar{Y}(j)\}^4$. This does not make any real difference when $\bar{Y}$ converges, but it allows us to generalize in the future if necessary. The existence of $\bar{Y}_\infty$ and S_∞ is for convenience. In fact, we only need $(\bar{Y}_N)$ to be bounded above (in norm), $\big(S_N(j,j)\big)_{N\geq 2J}$ to be bounded away from 0 for some j, and a finite 4th moment assumption for the same j. The proof of Theorem 1 reveals this. There is nothing special about the fourth moment in the preceding lemma, and the same argument works for any positive integer moment (with absolute values for odd integers). It is also interesting to note: even though $\sum_{i=1}^N (Y_i - \bar{Y})^2 \leq \sum_{i=1}^N Y_i^2$, it is not always the case that $\sum_{i=1}^N (Y_i - \bar{Y})^4 \leq \sum_{i=1}^N Y_i^4$.

Proposition 3 is based on [58], but its exact form does not appear there. We must perform one minor calculation before the result follows.

Proof. (of Proposition 3) The only thing that Theorem 5 in [58] does not do is give the asymptotic variances explicitly. These follow from Lemma 9: for $j, k = 1, \ldots, J$, $j \neq k$:

$$\mathrm{Var}(\hat{\bar{Y}}(j)) = \frac{N - N_j}{N \cdot N_j} S(j,j),\ \mathrm{Cov}(\hat{\bar{Y}}(j), \hat{\bar{Y}}(k)) = \frac{-S(j,k)}{N}$$ □

We have required Assumption B in order to work with the imputed potential outcomes and the permutation distributions. What we show next is that, under Assumption A, the imputed potential outcomes also satisfy Assumption A in probability. This is not strong enough for our purposes, and we make no use of it in the main text. However, we regard the result as being of some interest in its own right.

Lemma 15. *The imputed potential outcomes in FRT-2 and satisfy Assumption A in probability. That is, the sequences $(\bar{Y}^*)$ and (S^*) converge in probability, and $\max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2/N \xrightarrow{P} 0$.*

Lemma 4 is a much stronger statement than the last item of Lemma 11. It says we in fact have that $\max_{i,j}\{Y_i^*(j) - \bar{Y}^*(j)\}^2/N \to 0$ for all sequences of W. Not surprisingly, its proof is more involved.

Proof. It is immediate from (A.1) that the $\bar{Y}^*$ and $S^* = s^* 1_J 1_J^\top$ converge in probability because $\hat{\bar{Y}}(j)$ and $\hat{S}(j,j)$ do for all j, by Proposition 2. However, here we present a less streamlined version. We start with ANOVA identities

$$\bar{Y}_\bullet^{\mathrm{obs}} = \frac{1}{N}\sum_{i=1}^N Y_i^{\mathrm{obs}} = \sum_{j=1}^J \frac{N_j}{N}\hat{\bar{Y}}(j),$$

$$\sum_{i=1}^N (Y_i^{\mathrm{obs}} - \bar{Y}_\bullet^{\mathrm{obs}})^2 = \sum_{j=1}^J (N_j - 1)\hat{S}(j,j) + \sum_{j=1}^J N_j\{\hat{\bar{Y}}(j) - \bar{Y}_\bullet^{\mathrm{obs}}\}^2.$$

We recognize the second identity as the decomposition of total sum of squares into residual and treatment sum of squares, respectively. The means and variances of the imputed potential outcomes $Y_i^*(j) = Y_i^{\text{obs}} + z_j - z_{W_i}$ in FRT-2 are

$$\begin{aligned}
\bar{Y}^*(j) =& \frac{1}{N}\sum_{i=1}^{N}(Y_i^{\text{obs}} + z_j - z_{W_i}) = \sum_{k=1}^{J}\frac{N_k}{N}\hat{\bar{Y}}(k) + z_j - \bar{z},\\
S^*(j,k) =& \frac{1}{N-1}\sum_{i=1}^{N}\{Y_i^*(j) - \bar{Y}^*(j)\}\{Y_i^*(k) - \bar{Y}^*(k)\}\\
=& \frac{1}{N-1}\sum_{i=1}^{N}(Y_i^{\text{obs}} - z_{W_i} - \bar{Y}_\bullet^{\text{obs}} + \bar{z})^2 = s^*,\\
\text{where } s^* =& \sum_{j=1}^{J}\frac{N_j - 1}{N-1}\hat{S}(j,j) + \sum_{j=1}^{J}\frac{N_j}{N-1}\{\hat{\bar{Y}}(j) - z_j - \bar{Y}_\bullet^{\text{obs}} + \bar{z}\}^2.
\end{aligned}$$

This gives an equivalent formula for s^* as in (A.1). It is suboptimal because we did not develop notation to extend it to vector potential outcomes. From the above, it is clear that $\bar{Y}^*(j)$ and $S^*(j,k)$ converge in probability. We have

$$\bar{Y}^*(j) = z_j + \sum_{k=1}^{J}\frac{N_k}{N}\{\hat{\bar{Y}}(k) - z_k\} \xrightarrow{P} z_j + \sum_{k=1}^{J} p_k\{\bar{Y}(k) - z_k\}$$

and

$$\begin{aligned}
S^*(j,k) =& \sum_{j=1}^{J}\frac{N_j - 1}{N-1}\hat{S}(j,j) + \sum_{j=1}^{J}\frac{N_j}{N-1}\{\hat{\bar{Y}}(j) - z_j - \bar{Y}_\bullet^{\text{obs}} + \bar{z}\}^2\\
\xrightarrow{P}& \sum_{j=1}^{J} p_j\big[S(j,j) + \big\{\bar{Y}(j) - z_j - \sum_{k=1}^{J} p_k\bar{Y}(k) - \bar{z}\big\}^2\big]
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\max_{i,j}\frac{1}{N}\{Y_i^*(j) - \bar{Y}^*(j)\}^2 =& \max_i \frac{1}{N}\big(Y_i^{\text{obs}} - \bar{Y}_\bullet^{\text{obs}} - z_{W_i} + \bar{z}\big)^2\\
\le& \max_{i,j}\frac{1}{N}\big[Y_i(j) - z_j - \frac{1}{N}\sum_{k=1}^{J} N_k(\hat{\bar{Y}}(k) - z_k\}\big]^2\\
\le& \max_{i,j}\frac{1}{N}\{Y_i(j) - \bar{Y}(j)\}^2 + \max_{i,j}\frac{2}{N}\{Y_i(j) - \bar{Y}(j)\}\{\bar{Y}(j) - \tilde{z}_j\}\\
&+ \max_j \frac{1}{N}\{\bar{Y}(j) - \tilde{z}_j\}^2\\
\le& \max_{i,j}\frac{1}{N}\{Y_i(j) - \bar{Y}(j)\}^2 + \max_{i,j}\frac{2}{N}|Y_i(j) - \bar{Y}(j)| \cdot \max_j |\bar{Y}(j) - \tilde{z}_j|
\end{aligned}$$

which $\to 0$ in probability. For the equality we may drop the max over j because the expression does not involve j. In the second line we use that $Y_i^{\text{obs}} - z_{W_i} = Y_i(j) - z_j$ for some j. In the third line we use $\tilde{z}_j = z_j - \sum_{k=1}^{J} N_k\{\hat{\bar{Y}}(k) - z_k\}/N$ for convenience, add and subtract $\bar{Y}(j)$, FOIL, and use that the max of a sum is at most the sum of max's. In the fourth line we use Assumption A to show the first piece is zero and that $\{\bar{Y}(j) - \tilde{z}_j\}^2/N \xrightarrow{P} 0$ because $\tilde{z}_j$ converges in probability to a real number. The final step is to note $\max_j |\bar{Y}(j) - \tilde{z}_j|$ converges in probability to a real number, and if (a_i) is a sequence of real numbers, then $\lim_{N\to\infty} \max_{i=1,\ldots,N} a_i^2/N = 0$ implies $\lim_{N\to\infty} \max_{i=1,\ldots,N} |a_i|/N = 0$. □